Automatische Gewinnung enzymbezogener Informationen aus der wissenschaftlichen Primärliteratur

von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Carola Söhngen

aus Köln

1. Referent: Prof. Dr. Dietmar Schomburg

2. Referent: Prof. Dr. Dieter Jahn

eingereicht am: 29.06.2011

mündliche Prüfung (Disputation) am: 03.11.2011

Druckjahr 2011

Vorveröffentlichungen der Dissertation:

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen:

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D.: BRENDA, the enzyme information system in 2011. Nucleic Acids Research, (2010).

Tagungsbeiträge:

Söhngen, C., Scheer, M., Schomburg, I., Chang, A., Grote, A., Schomburg, D.: BRENDA text mining: new developments for obtaining enzyme-related disease information from the scientific literature. (Poster) F-19. European Conference on Computational Biology (ECCB), Ghent (Belgium) (2010).

Söhngen, C., Scheer, M., Schomburg, I., Chang, A., Grote, A., Schomburg, D.: BRENDA text mining: BRENDA text-mining enzyme-related disease information from the scientific literature(Poster))11. German Conference on Bioinformatics (GCB) 2010, Braunschweig

für Jochen in Liebe

"Central problem in philosophy. Relation of word to object . . . what is a word? Arbitrary sign. But we live in words. Our reality, among words not things. No such thing as a thing anyhow; a gestalt in the mind. Thingness . . . sense of substance. An illusion. Word is more real than the object it represents. Word doesn't represent reality. Word _is_ reality."

Philip K. Dick

TIME OUT OF JOINT (1959)

Inhaltsverzeichnis

Kurzzusammenfassung

Dynamische Umbrüche im Bereich der Lebenswissenschaften in den letzten Jahren, hier besonders das Entstehen der Systembiologie, ermöglichten einen globalen Blick der Forschung auf die Gesamtheit regulatorischer Prozesse innerhalb individueller Organismen. Dies ist die Folge der Einführung neuer Analyseverfahren, die nun im großen Maßstab Ergebnisse erzielen. Davon wird auch an den Schnittstellen der biomedizinischen Forschung zur Systembiologie profitiert, bei denen sich der Erkenntnisgewinn durch eine rasant wachsende Menge wissenschaftlicher Veröffentlichungen niederschlägt. Eine manuelle Annotation dieser Daten aus der Literatur zur Zusammenstellung themenspezifischer Datenbanken wird wegen ihrer Fülle zunehmend undurchführbar. Um mit Lösungsansätzen zur automatisierten Wissensgewinnung diesem Problem begegnen können, entwickelten sich neue Methoden in der Bioinformatik, die Verfahren aus der Computerlinguistik adaptieren und anwenden.

Die BRaunschweig ENzyme DAtabase (BRENDA) ist die weltweit größte im Internet verfügbare Sammlung von manuell aus wissenschaftlicher Primärliteratur gewonnenen Daten zu Enzymen. Seit einigen Jahren wird BRENDA durch Datenbanken ergänzt, deren Inhalt durch Methoden der automatisierten Wissensgewinnung generiert werden und zusammen das BRENDA Informationssystem bilden. Die vorliegende Dissertation stellt Lösungen vor, die eine Ausweitung des auf diese Weise gewonnen Datenfundus ermöglichen. Mit den in dieser Arbeit neu- und weiterentwickelten Methoden können in mehreren Schritten automatisiert relevante Referenzen gefunden werden, die Informationen über den Zusammenhang von Enzymen und Krankheiten enthalten. Zunächst werden Sätze und Titel aus Kurzzusammenfassungen, in denen Enzyme und Krankheiten gemeinsam auftreten mit hohem Erfolg (F_1 Maß 0,89) erfasst. Nachgelagert werden die semantischen Beziehungen von gemeinsam auftretenden Enzymen und Krankheiten anhand der Anwendung von Methoden des maschinellen Lernens klassifiziert.

Durch die Integration der so gewonnen Erkenntnisse in das BRENDA Informationssystem kann eine stetig erweiterte Auswahl von momentan über 500.000 referenzierten Publikationen, geordnet nach Enzymen und Krankheiten, für die sie relevant sind, abgefragt werden. Des Weiteren ist eine gezielte Suche nach Referenzen möglich, die wichtige Aussagen zur kausalen Verknüpfungen von Enzymen und Krankheiten, sowie der diagnostischen Verwendung und therapeutischen Implikation von Enzymen enthalten können ebenso wie nach Referenzen, die den aktuellen Stand der Forschung an einem Enzym und seiner Verbindung zu einer Krankheit widerspiegeln. Die automatisierte Erfassung von Zugangsnummern für Proteinsequenzen und -strukturen in biologischen Datenbanken aus Referenzen ermöglicht die Verknüpfung zu weiteren Informationsquellen und auch den Ringschluss vom Enzym zur Krankheit und zu aktuellen Forschungsansätzen neuer Therapien, über die organismusspezifische Sequenz und Struktur des Enzyms.

Abstract

The dynamic changes in life sciences led to the emergence of systems biology, which established a global view on the regulatory processes within organisms. The introduction of high-throughput analysis techniques went along with this development and a large number of results within a short period of time. The area of biomedical research benefits greatly and the number of scientific publications is increasing rapidly in this field. Manually curated life science databases are encountering their limitations and become incapable distributing information of all relevant articles. A fast growing branch within the field of bioinformatics adopted methods of automatic information retrieval originated by computational linguistics to meet this obstacle by the means text and data mining.

The BRaunschweig ENzyme DAtabase (BRENDA) is the most comprehensive online-available collection of manually annotated data on enzymes derived from scientific literature worldwide. Since a few years BRENDA is supplemented by databases which contain content that is generated by text mining approaches and thus build together the BRENDA knowledge system. This thesis introduces solutions to broaden the spectrum of data gained through text mining for the BRENDA knowledge system. In a multi-step procedure relevant publications containing information on enzymes and diseases are retrieved. First the abstracts of publications were successfully screened (F measure 0.89) for the co-occurrence of enzymes and diseases within one title or one sentence. This is followed by the classification of all co-occurring enzymes and diseases according to their semantic relation by the means of machine learning.

By the integration of the results into the BRENDA information system information on a constantly growing number of currently more than 500,000 relevant references with a enzyme and disease background are available. This collection is fully searchable and query results are clearly presented by arranging all retrieved references connected to an EC number and a disease. Furthermore a systematic search enables the retrieval of references that contain statements on the causal connection between enzymes and diseases and on enzymes that are used in the diagnostic process of a disease or enzymes which are drug targets or components of drugs to combat a certain disease. In addition references can be found which reflect the status of the ongoing research on enzymes which might be connected to diseases or important in future therapeutic approaches. The automatic extraction of life science database accession numbers for protein sequences and structures out of full-text references enables the ring-closure from the enzyme to the disease and the organism specific enzyme sequences addressed in the reference and the currently investigated research approaches towards new insights on disease mechanisms and therapies.

1. Einleitung

In den vergangenen Jahren haben die Lebenswissenschaften durch die Neu- und Weiterentwicklung von Untersuchungsmethoden, zum Beispiel in den Bereichen der Molekularbiologie, Biomedizin, Biochemie und Bioinformatik erhebliche Fortschritte gemacht. Insbesondere in dem integrativen Feld der Systembiologie wird versucht, weg von der isolierten Betrachtung einzelner Bausteine hin zur Aufklärung breiter Zusammenhänge zu wirken [1]. Die Hochdurchsatzanalyseverfahren, wie DNA-Sequenzierung oder Massenspektrometrie, haben auf den Gebieten der biologischen und medizinischen Forschung nicht nur große Mengen von Daten entstehen lassen, sondern auch eine stetig steigende Zahl von Publikationen, die dem Zweck folgen, den Erkenntnisgewinn zum öffentlichen Gut der wissenschaftlichen Gemeinschaft zu machen. Doch gerade dieser stetige Zuwachs führt dazu, dass das gesamte Spektrum der relevanten Veröffentlichungen, gerade in den sich schnell entwickelnden Teilgebieten, nicht mehr in angemessener Zeit durch Lektüre zu erfassen ist [2,3]. In Konsequenz hat eine Verfolgungsjagd begonnen, die Daten und Informationen, die ursprünglich als Ergebnisse der Analyseverfahren strukturiert vorlagen, nun in wissenschaftlichen Publikationen in natürlicher Sprache verfasst worden sind, wieder systematisch zu erfassen sowie möglichst verlässlich und aktuell für die schnelle und effiziente Suche vorzuhalten.

Um dieser Herausforderung zu begegnen, entstanden im Laufe der Zeit mehr und mehr wissenschaftliche Datenbanken (*1.3. Biologische und medizinische Wissenssammlungen*). Diese Datenbanken beinhalten kondensiertes und strukturiertes Wissen aus den jeweils adressierten Teilgebieten der Wissenschaft, das für unterschiedlichste Zwecke eingesetzt werden kann. Datenbanken stoßen aber überall dort an die Grenze der Vervollständigung, wo die Arbeit per Hand von Experten für die Extraktion wissenschaftlicher Daten aus Primärliteratur notwendig ist. Diese Art der Recherche und manuellen Annotation relevanter Referenzen ist immer noch unübertroffener Garant für die Qualität des Inhalts einer Datenbank, gleichzeitig aber auch mit erheblichem Zeitaufwand und Kosten verbunden. Seit Rechnersysteme eine Leistungsfähigkeit erreicht haben, die komplexe Prozessierungsschritte in angemessener Zeit bewältigen, wie sie für die automatische Wissensgewinnung aus Texten notwendig sind, ist die zunehmende Automatisierung dieser Tätigkeiten möglich geworden. Neben der reinen Rechenleistung (Hardware) sind aber auch Strategien und Algorithmen und deren Umsetzung (Software) notwendig für die Bewältigung von großen Mengen unstrukturierter Daten in Textform.

1.1. Automatische Informationsgewinnung

Die Disziplin der automatischen Informationsgewinnung aus Daten, und hier insbesondere Texten (Text Mining), entwickelte sich aus der Schnittmenge der Anforderungen an die Qualität der gewonnenen Resultate und der gleichzeitigen Optimierung der Effizienz der Verarbeitung. Die Ziele und Namensgebungen für die automatische Informationsgewinnung aus Text und anderen Daten sind seit dem Entstehen dieser Teildisziplin aus der Linguistik und der (Bio-)Informatik, je nach Zweck und Aufgabenstellung, unterschiedlich definiert worden [4].

Aus dem Grundtenor können aber zwei Kernkonzepte diesem Gebiet zugeordnet werden [3,5-8], die in dieser Arbeit im Zentrum stehen:

- **Identifikation von Entitäten:**
  Die Identifizierung beliebiger relevanter Objekte, Größen, Einheiten oder Eigenschaften.
- **Klassifizierung von Entitätsbeziehungen:**
  Das Aufzeigen und Einordnen von bestenfalls neuen und unter Umständen nicht erwarteten Zusammenhängen.

Die Entität

Bevor detaillierter auf die Konzepte der Identifikation von Entitäten und deren Relationsklassifizierung eingegangen wird, wird zunächst die Bedeutung des Begriffs *Entität* näher definiert. Der Begriff Entität[1] findet sowohl in der Philosophie als auch in der Linguistik Verwendung. In der Ontologielehre der Philosophie bezeichnet es das Dasein eines Dinges, und in der Linguistik steht *Entität* für eine beliebige Größe, Einheit oder Eigenschaft [10]. Beide Definitionen sind für die Verwendung des Begriffes in dieser vorliegenden Arbeit relevant. Eine Entität kann eine Stadt, eine Firma oder ein Protein sein. Um nach einer wie auch immer gearteten Entität suchen zu können, muss sie benennbar und abgrenzbar von einer anderen Entität sein; um sie zu finden, muss die Entität im Suchbereich präsent sein.

In der biomedizinischen Informationsgewinnung aus Daten- und Texten ist auch der Begriff *biologische Entität* (biological entity) zu finden [10-12], der Entitäten wie Proteine, Organismen oder Gennamen umfassen kann. Biologische Entitäten haben

[1] von mittellateinisch entitas - das Wesen des Dinges [9]

durch Nomenklaturvorgaben und einem stetigen Wandel durch Neuentdeckung und Verwurf einen besonderen Status, der sie von der Benennung anderer Entitäten abhebt [12,14]. Die Entität „Alkoholdehydrogenase“ umfasst ein Enzym. Es ist der *Enzyme Commission* [15] (EC) Nummer 1.1.1.1 zugeordnet. Es ist unerheblich, ob diese Entität durch andere sie-bezeichnende Begriffe (Synonyme) wie „ADH“, oder durch einen Namen in einer anderen Sprache (eng. alcohol dehydrogenase) benannt oder mit einem Satz „Das Enzym, das die chemische Umsetzung von Alkoholen zu den entsprechenden Aldehyden katalysiert“ umschrieben wird. Alle Arten der Benennung und Umschreibung haben gemein, dass sie sich immer auf die eine Entität der „Alkoholdehydrogenase“ beziehen.

Entitäten suchen und finden

Grundlegend für die Entscheidung, ob Texte einen relevanten Inhalt haben, ist die Präsenz der Objekte, die im Fokus des Interesses stehen. Deswegen ist eine Identifizierung von Entitäten meist ein erster Schritt um Textinhalte zu erfassen [12,13,16].

Die Identifizierung von biologischen Entitäten ist aufgrund des Charakters des biomedizinischen Vokabulars, welches durch Nomenklaturvorgaben und deren Missachtung hoch komplex ist sowie durch die fortlaufende Forschung und Neuentdeckungen schnell wächst, eine anspruchsvolle Aufgabe mit spezifischen Anforderungen [11,12,14]. Es gibt unterschiedliche Ansätze, um Entitäten in Texten zu identifizieren. Die Kenntnis von Namen und synonymen Bezeichnungen ermöglicht die Erstellung eines Wörterbuchs für die einzelnen Entitäten und erlaubt so eine systematische Suche in Texten. Wörterbücher zu biologischen Entitäten können aus den entsprechenden Datenbanken, Ontologien und Thesauri extrahiert werden, sofern diese vorhanden und frei zugänglich sind.

Die Schwierigkeiten bei der Suche anhand von entsprechenden Wörterbüchern sind nicht zuletzt die zahlreichen Namen, die sich gleich lautend auf mehr als eine Entität beziehen, sogenannte Homonyme. Homonyme sind meist nur aus dem Zusammenhang heraus eindeutig, wie Maus: Der Satz „Die Maus ist auf dem Tisch“ ist mehrdeutig und könnte beispielsweise eine Computermaus oder das Tier Maus meinen. Akronyme und Abkürzungen, die in der wissenschaftlichen Literatur häufig Verwendung finden, können ebenso die Identifizierung einer Entität erschweren. Das Akronym „AAA“ wird unter anderem für das Enzym Aryl-acylamidase (EC 3.5.1.13) verwendet. In (bio-)medizinischer Literatur könnte es aber auch für „abdominales Aortenaneurysma“, das Tripeptid „L-Alanyl-L-alanyl-L-alanine“ oder „American Ambulance Association“

stehen. Wörterbücher, die der Identifizierung von Entitäten dienen, sollten dahingehend einer Optimierung unterzogen werden bei der gegebenenfalls diese Begriffe und Abkürzungen gelöscht werden. Sofern gerade ein solcher Name sehr häufig für die Entität verwendet wird, und er deshalb in dem Wörterbuch verbleiben muss, sollten weitere filternde Schritte der Disambiguierung angeschlossen werden. Folgen die Zeichenketten, die den Namen einer Entität bilden, ableitbaren Mustern, so ermöglicht das eine regelbasierte Suche nach Entitäten, bei der auf die Anwendung eines Wörterbuchs verzichtet werden könnte.

Klassifizierung von Entitätsbeziehungen

Die Identifizierung von relevanten Entitäten ist nur ein erster Schritt, um Aussagen und deren Bedeutung in Texten zu erfassen und zu differenzieren [11,12]. Ein konsequenter Schritt im Anschluss an die Identifizierung zweier Entitäten ist die Betrachtung ihrer Distanz zueinander. Das gemeinsame Auftreten zweier Entitäten innerhalb einer definierten sprachlichen Einheit, wie einer Phrase oder einem Satz, wird in der Linguistik als Kookkurrenz bezeichnet [17]. Die Kookkurrenz von Entitäten in einem Satz bedingt in der überwiegenden Zahl der Fälle eine semantische Verknüpfung dieser [12,18]. Dennoch ist die Art der semantischen Beziehungen, die dadurch bestehen könnten, nicht immer eindeutig davon abzuleiten. Deswegen ist eine nachgeschaltete Verarbeitung notwendig um weitere differenziertere Einteilungen vornehmen zu können. Bei sehr großen Kollektionen von Texten (Textkörpern), können selbst nach dem Ausschluss aller Sätze, in denen keine Kookkurrenz festgestellt wurde, noch so große Textmengen zur weiteren Analyse verbleiben, dass auch diese nicht ohne erheblichen Aufwand bei einem Verzicht auf Automatisierung ausgewertet werden können. Anhand von kleinen Teilmengen von Material, das für die entsprechende Fragestellungen ausgewertet vorliegt, ist es möglich, Methoden des maschinellen Lernens anzuwenden, die unter anderem die automatische Klassifizierungen von semantischen Beziehungen ermöglichen. Eine dieser Methoden ist die Klassifizierung durch eine *Support Vector Machine* (SVM). Das mathematische Konzept der SVM beruht auf einer Trennung von Objekten unbekannter Klasseneinordnung, durch vorherige Analyse von Objekten bekannter Klasseneinordnung [19].

1.2. Enzyme, ihre Funktion und Fehlfunktion

Einzellige Organismen bestehen bereits aus komplexen Systemen verschiedenster Regelkreise und Vorgänge, die für das Leben notwendig sind. Die Komplexität vervielfacht sich noch einmal für höhere Lebewesen, wie Säugetiere, die aus einer

Vielzahl von Zelltypen und Gewebearten bestehen; jede für sich ist zum Teil hochspezialisiert und beinhaltet vielschichtig regulierte biochemische Vorgänge. Den Enzymen kommt in dem metabolischen Gefüge von Organismen eine besondere Rolle zu. Sie katalysieren nahezu alle biochemischen Reaktionen in der belebten Natur [20]. Die katalysierten Reaktionen lassen sich verschiedenen Reaktionstypen zuordnen (siehe Tabelle 1.1). Diese Reaktionstypen bilden die sechs Hauptklassen des von der *International Union of Biochemistry* (IUBMB) [15] entwickelten Enzym-Nomenklatur Systems. Bei einer offiziellen Benennung durch die Enzymkommission der IUBMB wird jedem Enzym ein empfohlener Name (recommended name), ein systematischer Name (systematic name) und eine vierstellige *Enzyme Commission* (EC) Nummer zugeordnet. Eine EC Nummer kann Enzyme aus unterschiedlichen Organismen umfassen.

Wie bei jedem System, das derart verflochten ist, wie der Metabolismus von Organismen, so bleiben Störungen meist nicht ohne Folgen. Da die katalytische Funktion von Enzymen eine derart essenzielle Rolle im metabolischen Gefüge spielt, können deren Beeinträchtigung fatale Auswirkungen haben (siehe Tabelle 1.1). Wenn eine Mutationen in dem kodierenden Gen eines Enzyms eine Veränderungen in der Aminosäuresequenz bedingt, kann die Enzymfunktion komplett aufgehoben sein. Dies tritt beispielsweise bei der *Phenylketonurie* (Tabelle 1.1) auf, eine der häufigsten angeborenen Stoffwechselerkrankungen [21], für die die Fehlfaltung des Enzyms *Phenylalanin 4-Monooxygenase* (EC 1.14.16.1) verantwortlich ist. Einige pathologische Vorgänge können aber auch durch voll funktionstüchtige Enzyme verursacht werden, wie bei der Nervengiftfreisetzung durch *Clostridium botulinum*, einem pathogenen Bakterium, dessen Enzym *Bontoxilysin* (EC 3.4.24.69) den Botulismus verursacht.

Tabelle 1.1: Die sechs Enzymklassen, entsprechend der IUBMB Klassifikation. Den Enzymklassen sind die Namen der Enzymklasse sowie die allgemein katalysierten Reaktionstypen zugeordnet (nach Voet Biochemistry [20]). Für jede Enzymklasse ist ein Beispiel einer Krankheit angegeben, deren Pathogenese maßgeblich mit einem Enzym aus dieser Klasse verknüpft ist. Das jeweilige Enzym ist zusammen mit seiner Enzyme Commission (EC) Nummer angegeben. Die vollständige EC-Nummer eines Enzyms besteht aus vier Ziffern, die durch Punkte getrennt werden. Die Ziffern bezeichnen die Klasse, Subklasse, Sub-Subklasse und die laufende Nummer innerhalb der Sub-Subklasse.

Enzym Klasse	Name	Katalysierte Reaktionstypen	Beispiele für Krankheiten und Enzyme, die bei der Pathogenese eine zentrale Rolle spielen
1	Oxidoreduktasen	Redoxreaktionen; Elektronentransfer zwischen zwei Stoffen	Krankheit: Phenylketonurie Ursache: Mutation im kodierenden Gen des Enzyms Phenylalanin 4-Monooxygenase (EC 1.14.16.1)
2	Transferasen	Transfer von chemischen Gruppen von einem Donor auf einen Akzeptor	Krankheit: Cholera Ursache: Exotoxin aus *Vibrio cholerae* Enzym: NAD+-diphthamide ADP-ribosyltransferase (EC 2.4.2.36)
3	Hydrolasen	Hydrolytische Spaltung (Abgang von H_20)	Krankheit: Botulismus Ursache: Nervengiftfreisetzung durch *Clostridium botulinum* Enzym: Bontoxilysin EC 3.4.24.69
4	Lyasen	Spaltung mit Abgang von chemischen Stoffen (nicht hydrolytisch)	Krankheit: Hereditäre Fruktoseintoleranz Ursache: Mutation im kodierenden Gen des Enzyms Enzym: Fruktose-bisphosphat Aldolase (EC 4.1.2.13)
5	Isomerasen	Veränderung des Substrates ohne Änderung der Summenformel (intramolekularer Transfer)	Krankheit: Früh-/Fehlgeburt Ursache: Mutation im kodierenden Gen des Enzyms Enzym: Phosphoglucomutase (EC 5.4.2.2)
6	Ligasen	Verbindung zweier Moleküle auf Kosten einer energiereichen Phosphatbindung (ATP Spaltung)	Krankheit: Multipler Carboxylase-Mangel Ursache: Mutation im kodierenden Gen des Enzyms Enzym: Holocarboxylase Synthetase (EC 6.3.4.10)

Ebenso beruhen viele pharmakologische Ansätze zur Therapie von Krankheiten auf der Beeinflussung mittels Hemmung oder Anregung eines Enzyms. Ein Enzym kann auch als Bestandteil eines Medikaments seine Wirkung im Organismus entfalten. Beispielsweise wird das gleiche Enzym, dass mit dem Botulismus eine schwere Erkrankung auslöst, zur Therapie bei *Spasmus hemifacialis* eingesetzt [22].

Enzyme sind zudem Teil der Abklärung eines pathologischen Zustands (Diagnose) und der Kontrollen im Rahmen von Vorsorgeuntersuchungen und Screeningverfahren. Es werden quantitative und qualitative Bestimmungen, Nachweise von organspezifischen (Iso-)Enzymen oder sogar krankheitsspezifische Muster für ganze Enzymgruppen bestimmt [23]. Beispiele für diagnostische Laboruntersuchungen sind die Ermittlung der Aktivität von Enzymen: Eine Abweichung der Aktivität der *Phosphopyruvat Hydratase* (EC 4.2.1.11) oder des *prostataspezifisches Antigens* (EC 3.4.21.77) von Normalwerten gilt als Hinweis für das Vorliegen von unterschiedlichen Neoplasien. Diese Enzyme werden deshalb auch zu den *Tumormarkern* gezählt. Die *Kreatinkinase* (EC 2.7.3.2) gilt als ein Leitenzym für die Diagnose von Erkrankungen in Zusammenhang mit der Herz- und Skelettmuskulatur [23].

1.3. Biologische und medizinische Wissenssammlungen

Um die Erkenntnisse in den Lebenswissenschaften bedarfsgerecht vorzuhalten und zu verbreiten, entstanden eine Vielzahl von Datensammlungen und Datenbanken, die elektronisch (und zum Teil gedruckt) zur Verfügung stehen. Die Themenschwerpunkte und die Inhaltsausformung richten sich nach dem abgedeckten Fachgebiet und reichen von „sehr speziell und eng umgrenzt“ bis hin zu „möglichst allumfassend“. Die Einträge in unterschiedlichen wissenschaftlichen Datenbanken besitzen ein uneinheitliches Format und können die verschiedensten Datenfelder enthalten. Um einen bestimmten Eintrag zu referenzieren haben Datenbanken einen eindeutigen Identifikator. Dieser Identifikator ist meist eine alphanumerische Zeichenfolge und wird auch Zugangsnummer (eng. accession number) genannt. In Texten mit naturwissenschaftlichen Hintergrund wird zur eindeutigen Bestimmung einer beschriebenen Entität, z.B. eines Proteins, häufig begleitend die entsprechende Zugangsnummer einer Datenbank angegeben.

Gesammelte Literaturreferenzen und deren Begriffswelten

Die *PubMed* [24] Datenbank gehört mittlerweile zu einer der wichtigsten Verzeichnisse wissenschaftlicher Primärliteratur der Lebenswissenschaften und deren überschneidenden Bereichen. PubMed gehört zu den Datenressourcen des *National Center for Biotechnology Information* (NCBI) [24]. Sie umfasst über 20 Millionen Einträge zu Referenzen, von denen mehr als 11 Millionen mit einer Kurzzusammenfassung und ebenso viele mit einem Verweis zu der Volltextversion der Referenz versehen sind [24]. Die frei verfügbaren Kurzzusammenfassungen sind mit wichtigen Fachbegriffen indiziert und dienen neben der Literaturrecherche auch vermehrt Ansätzen aus dem Bereich der automatisierten Wissensgewinnung [25]. Diese Indexierung erfolgt manuell anhand von Fachvokabular aus den *Medical Subject Headings* (MeSH) [26]. MeSH ist ein hierarchisch organisierter Thesaurus dessen kontrolliertes Vokabular aus dem Bereich der Lebenswissenschaften und der Medizin stammt. Zur Zeit (Juni 2011) besteht die MeSH Kollektion aus ca. 26.000[2] Begriffen. Die Sammlung des Fachvokabulars wird durch die *United States National Library of Medicine* (NLM) [27] gepflegt. Die unterschiedlichen Begriffssammlungen, sofern sie semantisch strukturiert vorliegen auch Ontologien genannt, finden mittlerweile häufig Verwendung in Text Mining Ansätzen aus dem Bereich der Lebenswissenschaften. Eine dieser Ontologien ist das *Unified Medical Language System* ® (UMLS) [28] von der NLM, welches biomedizinische Begriffe aus medizinischen Wörterbüchern und

[2] Fact Sheet Medical Subject Headings 2011

Internetdatenbanken zusammenfasst, angleicht und miteinander in Relation setzt [29]. Obwohl sehr umfangreich (über 5 Mio. Konzeptnamen), ist es für den akademischen Gebrauch, durch die vielen kommerziellen Einrichtungen, die Rechte auf Teilvokabular inne haben, nicht ohne weitere Auflagen frei verfügbar und seine Verwendung bedingt einen jährlichen Bericht an die NLM.

Automatische generierte Sammlungen

Die manuelle Erfassung und Annotation von Daten ist inzwischen in einigen Datenbanken durch automatische Verfahren der Wissensgewinnung ergänzt worden. Einige Informationsplattformen und Datenbanken werden sogar vollständig durch die Entwicklung und Anwendung dieser automatischen Verfahren erstellt.

Die Plattform *GoPubMed* [30] verwendet Ontologien, wie *Gene Ontology* [31] als Indexierungsgrundlage, um PubMed Referenzen weitergehend zu unterteilen und so die Literatursuche zu beschleunigen. Die Vernetzung der Einträge mit anderen Quellen im World Wide Web soll so optimiert werden, dass Fragestellungen nach semantischen Verknüpfungen beantwortet werden können [32]. Dieser Aufbau einer vernetzten Strukturierung des Wissens wird semantisches Netz (eng. semantic web) genannt. Eine weitere Literatursuchmaschine dieser Art ist *iHOP* [33]. Die iHop Quelle sind PubMed Kurzzusammenfassungen, die automatisch mit Gen- und Proteinnamen sowie MeSH Begriffen indexiert werden. Durch Assoziation mit bestimmten Verben und Informationen zu deren physikalischen Interaktionen in regulatorischen Netzwerken soll eine semantische Einteilung erreicht werden. Die Suchmaschine *MedEvi* [34] bietet bei Mehrbegriffsuchen neben der gefundenen Referenz, die gefundenen Begriffe und deren semantische Relation hervorgehoben in einem „Beweissatz“. Es existieren auch Verfahren, die automatisch Informationen zu eng umgrenzten Gebieten und Fragestellungen extrahieren, wie z.B. der ausschließlichen Suche nach Mutationen von Genen der humanen Kinasen [35]. Die Webplattform *EnzyMiner* [36] bietet Informationen zu 22 Enzymen, für die automatisiert Erkenntnisse über Mutationen auf Proteinebene zusammengetragen wurden, die teilweise auch in Verbindung mit Krankheiten stehen können. Informationen zu weiteren Enzymen müssen per Mailformular angefordert werden und sind somit nicht sofort abrufbar.

Sequenzdatenbanken

Neben den Verzeichnissen für Literatur oder Vokabular gibt es ferner Datenbanken, die sich auf die Sammlung von Informationen zum Aufbau der biochemischen Makromoleküle wie Proteinen fokussieren. Sequenzdatenbanken enthalten Einträge zu

biologischen Verbindungen, die sich in ihrer Primärstrukur als Sequenzabfolgen einzelner Untereinheiten abbilden lassen. Sie enthalten darüber hinaus auch meist die Information über den Quellorganismus, aus dem die Sequenz isoliert worden ist. Man unterscheidet Nukleinsäure-, Protein- und Genomdatenbanken.

Zu den bedeutendsten Nukleinsäuresequenzdatenbanken gehören die Sammlungen des *European Molecular Biology Laboratory's European Bioinformatics Institute* (EMBL-EBI) [37], die *DNA Databank of Japan* (DDBJ) [38] und die *GenBank* [39] des NCBI. Die Proteindatenbank *UniProt* [40] enthält u.a. Informationen über die Proteinaufbau und deren Funktion. Die *UniProt* Datenbank bietet auch eine breite Informationsfülle durch Kreuzverweise auf bis zu 120 weitere Datenbanken aus allen relevanten Bereichen der Lebenswissenschaft, z.B. zu den 3D Strukturen von Proteinen in der *Protein Data Bank* (PDB) [41] oder zu den einzelnen Domänen, Strukturmotiven und funktionellen Gruppen von Proteinen in *PROSITE* [42]. Die katalytischen Mechanismen eines Enzyms werden oft erst durch das Wissen über seine dreidimensionale Struktur aufgeklärt. Deswegen sind die Verweise zu Datenbanken mit dreidimensionalen Strukturen von Proteinen, gerade im Zusammenhang mit Enzymen, eine wichtige Erkenntnisquelle.

Das Enzyminformationssystem BRENDA

Die *BRaunschweig ENzyme DAtabase* (BRENDA) [43] ist spezialisiert auf Informationen zu Enzymen und dafür die weltweit größte Ressource. BRENDA wird durch die manuelle Annotation aus Primärliteratur seit 1987 gepflegt [44] und bietet Auskunft über die IUBMB konforme Klassifikation und Nomenklatur sowie eine kategorisierende Erfassung von über 4.740 EC Nummern (Stand Juli 2011) und zahlreiche andere Aspekte aus über 100.000 Referenzen. Es finden sich Informationen, u. a. über die katalysierten Reaktionen und die Spezifität eines Enzyms, die funktionellen Parameter und seine kinetischen Werte, die Lokalisation des Enzyms im Organismus und seine Struktur sowie über Mutationen in kodierenden Genen von Enzymen.

Die manuell extrahierten Inhalte in BRENDA werden seit 2006 auch durch zwei Datenbanken ergänzt, die deren Informationen durch Text Mining Ansätze gewonnen werden: AMENDA (Automatic Mining of ENzyme DAta) und FRENDA (Full Reference ENzyme DAta) [45]. Die Referenzen in FRENDA enthalten organismenspezifische Enzyminformationen und AMENDA ist eine Teilmenge von FRENDA. Die in AMENDA enthaltenen Referenzen sind neben ihrer Aufschlüsselung nach Organismen auch nach Angaben zu der Lokalisation des Enzyms sowie der

enthaltenden Gewebeart erfasst. Alle automatisiert gewonnen Einträge können wahlweise zusätzlich zu den manuell annotierten Informationen in BRENDA angezeigt werden und bilden auf diese Weise ein umfassendes Informationssystem [43,46].

1.4. Aufgabenstellung

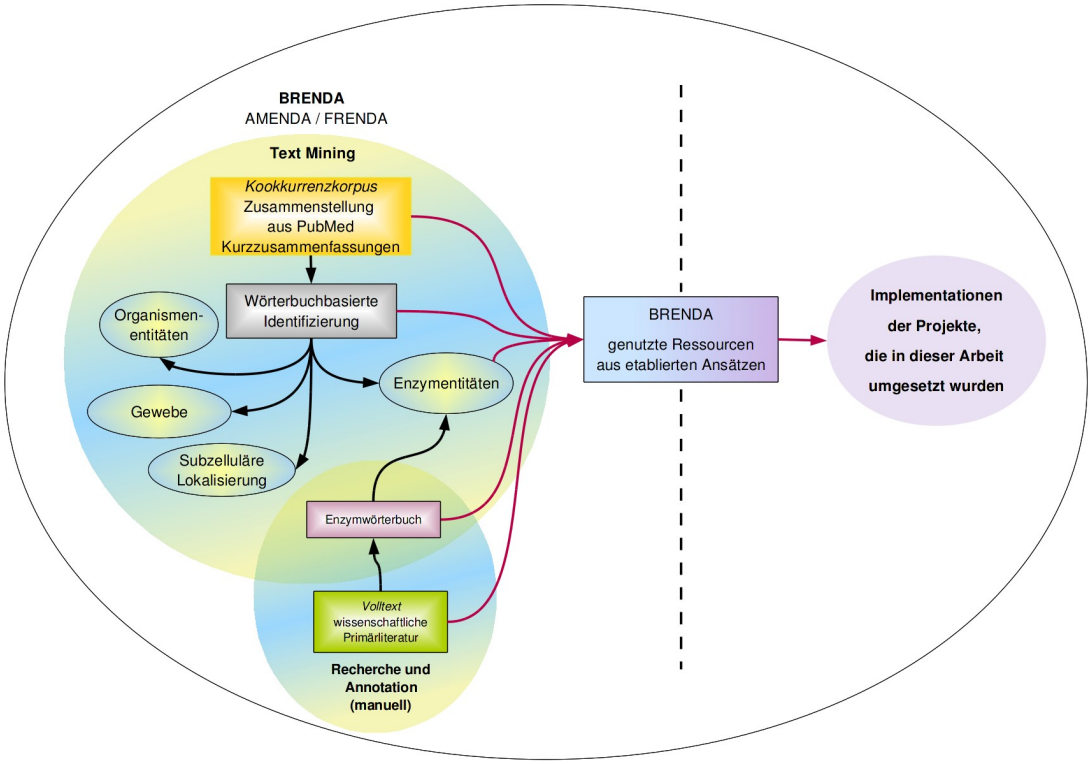
Ziel dieser Arbeit war die automatisierte Erfassung und Analyse von Literatur, deren Inhalt für eine weitere Auswertung zu Aspekten und Fragestellungen im Zusammenhang mit Enzymen relevant ist. Es sollten dazu bestehende Ansätze zur automatisierten Textauswertung des BRENDA Informationssystems als Ausgangspunkt genutzt werden und an Schnittstellen Neuentwicklungen sowie Erweiterungen integriert werden. Die Teilaufgaben bezogen sich insgesamt auf die Gewinnung enzymbezogener Informationen aus wissenschaftlicher Primärliteratur und stellten sich im Einzelnen wie folgt dar:

1. Es sollten anhand der Auswertung der in PubMed enthaltenen Titel und Kurzzusammenfassungen Referenzen erfasst werden, die Informationen zu Enzymen im Zusammenhang mit pathologischen Vorgängen im Organismus enthalten.
2. Die Titel und Kurzzusammenfassungen der so erfassten Referenzen sollten weiterhin klassifiziert werden, um Aufschluss darüber zu bekommen, in welchem Zusammenhang das beschriebene Enzym und die gefundene Krankheit stehen. Dazu sollte eine kategorisierende Einteilung definiert werden, die Aspekte beschreibt, unter denen Enzym und Krankheit in Verbindung stehen können. Darunter fallen die Verknüpfungen der Enzymfunktion oder deren Störung als Ursache der Krankheit sowie die Beeinflussung der Enzymfunktion durch Vorgänge im Rahmen der Erkrankung. Des Weiteren die Verbindung der Krankheit durch diagnostische Anwendung des Enzyms, zu deren Feststellung und Verlaufskontrolle oder die Beschreibung therapeutischer Maßnahmen gegen die Erkrankung, die mit dem Enzym als Wirkstoff in der Arznei oder als Wirkziel im Organismus verbunden sind.
3. Anhand der Entwicklung einer Anwendung, die Methoden des maschinellen Lernens nutzt, sollten die im ersten Schritt gefundenen Referenzen, die eine Krankheit und ein Enzym behandeln, durch eine automatische Klassifizierungsroutine den definierten inhaltlichen Schwerpunkten eingeordnet werden.
4. Um einen Ringschluss der eindeutigen Zuordnung der Referenz zu Enzym und Organismus zu ermöglichen, sollten in einer Volltextauswertung der gesamten Referenz, die darin enthaltenen Zugangsnummern zu Sequenz- und Strukturdatenbanken extrahiert werden.

2. Daten, Algorithmen und Methoden

Im folgenden Kapitel sollen die entwickelten Methoden zur automatischen Identifizierung, Einordnung und weitergehenden Datenextraktion beschrieben werden. Die Aufgabenstellung und die angedachte Nutzung der entwickelten Elemente im BRENDA Informationssystem bedingen eine Verzahnung der genutzten Materialien und Methoden mit bereits bestehenden Ressourcen. Zum Teil wurde deswegen für diese Arbeit auf etablierte Ressourcen zurück gegriffen, es wurden Ansätze bestehender Methoden adaptiert, die der manuellen Recherche und Annotation für BRENDA und den durch Text Mining Methoden generierten Datenbanken AMENDA und FRENDA [45] entspringen. Eine graphische Übersicht über diese Ressourcen befindet sich in Abbildung 2.1.

Abbildung 2.1: In den blau-gelben Bereichen sind die relevanten Ressourcen der BRENDA, AMENDA und FRENDA Datenbanken, die auch in Teilen für deren eigene Erstellung genutzt werden, dargestellt. Die roten Pfeile verdeutlichen, welche Ressourcen für die Umsetzung der Ansätze dieser Arbeit einbezogen wurden.



2.1. Auswahl und Zusammenstellung von Sprachdaten

Den Beginn bildet eine Betrachtung der als Ausgangsmaterial verwendeten Textkörper und Wörterbücher in dem Abschnitt *2.1. Auswahl und Zusammenstellung von Sprachdaten* und deren Aufbereitung im Abschnitt *2.2. Textaufbereitung und Datenstrukturen*.

2.1.1. Textkorpus

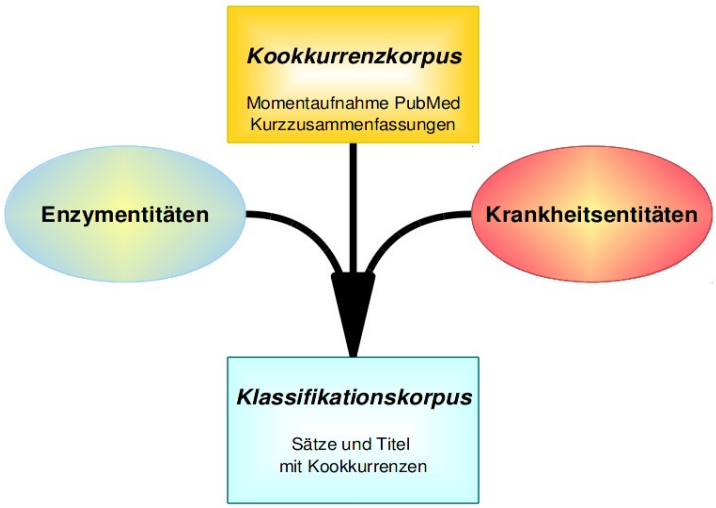
Der Begriff Textkorpus (syn. Textkörper) bezeichnet eine Sammlung von Sprachdaten. Die Sprachdaten können schriftliche Texte in jeder Form, wie Bücher, Artikel in Zeitschriften oder Protokolle, umfassen. Die Größe, Art und Zusammensetzung eines Textkörpers werden durch die jeweilige Fragestellung bzw. Verwendungszweck beeinflusst. Die in der Arbeit verwendeten Sprachdaten liegen ausschließlich in elektronischer Form und in englischer Sprache vor. Als Materialgrundlage wurden unterschiedliche Textkörper zusammengestellt und verarbeitet, die folgendermaßen aufgebaut und benannt wurden:

Volltextkorpus: Die Zusammenstellung umfasst wissenschaftliche Artikel, deren Volltext aus Portable Document Format[3] (PDF) Dateien extrahiert wurde. In diesem *Volltextkorpus* wurde eine Suche nach Zugangsnummern von verschiedenen biologischen Datenbanken vorgenommen. Der Inhalt dieser Artikel war vorher im Rahmen manueller Recherchen von Experten als relevant im Hinblick auf eine manuelle Auswertung und Aufnahme in BRENDA, für ein oder mehrere Enzyme, befunden worden.

Kookkurrenzkorpus: Dieser Korpus beinhaltet eine Sammlung aller in der *PubMed* [24] Datenbank enthaltenen Kurzzusammenfassungen (Abbildung 2.2). Der *Kookkurrenzkorpus* dient der Identifizierung der Krankheitsentitäten sowie der Kookkurrenzsuche von Krankheits- und Enzymentitäten. Er wird bei jedem Aktualisierungszyklus von BRENDA neu gebildet. Er stand als verwendete, nicht selbst gebildete Ressource für diese Arbeit zur Verfügung.

[3] Das plattformunabhängige Portable Document Format wurde vom Unternehmen Adobe Systems 1993 entwickelt und veröffentlicht. Seit 1997 sind der *International Organization for Standardization* (ISO) für dieses Dateiformat definiert, deren Spezifikationen gegen Gebühr unter www.iso.org verfügbar sind.

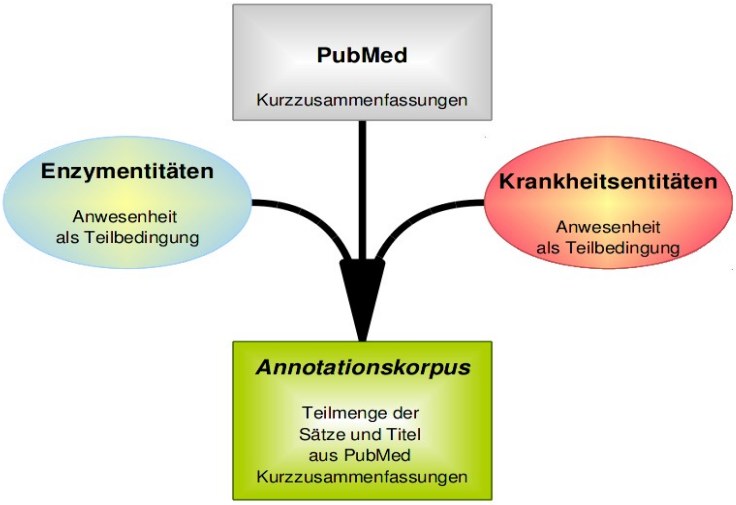
Abbildung 2.2: Eine schematische Darstellung der Zusammensetzung des Kookkurrenzkorpus und des Klassifikationskorpus und ihrer Beziehung zueinander. Der Kookkurrenzkorpus besteht aus allen Titeln und Sätzen der PubMed Kurzzusammenfassungen, die zum Zeitpunkt seiner Bildung in PubMed enthalten waren und wurde von der BRENDA Aktualisierungsroutine (Abbildung 2.1) als Material übernommen. Der Klassifikationskorpus ist eine Teilmenge des Kookkurrenzkorpus. Er besteht aus allen Titeln und Sätzen des Kookkurrenzkorpus, in denen mindestens einmal eine Enzymentität und eine Krankheitsentität gemeinsam durch die Entitätsidentifizierung festgehalten wurden.



Klassifikationskorpus: Zur Klassifizierung der semantischen Relationen von Enzym- und Krankheitsentitäten wurde eine Teilmenge des Kookkurrenzkorpus eingesetzt. Der *Klassifikationskorpus* besteht aus allen Titeln und Sätzen der Kurzzusammenfassungen, in denen eine Kookkurrenz von mindestens einer Enzymentität und einer Krankheitsentität gefunden wurde (Abbildung 2.2).

Annotationskorpus: Zum Zweck der Validierung der erzielten Ergebnisse der Kookkurrenzsuche und als Ressource für Trainings und Testdaten für die weitere Klassifizierung dieser Ergebnisse, wurde eine Kollektion aus 5031 Titeln und Sätzen aus den Kurzzusammenfassungen der PubMed Datenbank zusammengestellt. Um den *Annotationskorpus* mit potentiell relevanten Einträgen anzureichern, fand die Auswahl einzelner Anteile unter Vorbedingungen statt. Die Vorbedingungen für die Auswahl, wurden zum Teil an das Vorhandensein von Enzym- bzw. Krankheitsentitäten gebunden (Abbildung 2.3).

Abbildung 2.3: Skizze zur Auswahl der Titel und Sätze, die den Annotationskorpus bilden. Der Annotationskorpus wurde einmalig zusammengestellt und ist im Gegensatz zu den anderen Textkörpern statisch. Er kann erweitert werden, aber wird nicht vor jedem neuen Verarbeitungsdurchlauf neu gebildet.



Die Information für die Überprüfung der Vorbedingungen kam aus unterschiedlichen Quellen. Zum einen wurden die Ergebnisse der Suche nach Enzymentitäten (Abbildung 2.1) im Rahmen der bereits vor dieser Arbeit etablierten BRENDA Aktualisierungsroutine [46] übernommen und verwendet, zum anderen fanden die Ergebnisse, der in dieser Arbeit entwickelten Suche nach Krankheitsentitäten Berücksichtigung. So wurde der *Annotationskorpus* unter folgenden Bedingungen aus drei Anteilen zusammengestellt:

- Eine Menge von ca. 2.500 Sätzen, beziehungsweise Titeln wurde randomisiert aus der Gesamtmenge aller Sätze und Titel der PubMed Kurzzusammenfassungen gewählt, die eine Enzymentität enthalten.
- Ein Anteil von ca. 2.000 Titeln und Sätzen wurde randomisiert aus der Gesamtmenge aller Sätze und Titel der PubMed Kurzzusammenfassungen ausgewählt, wenn mindestens eine Enzymentität und eine Krankheitsentität als Kookkurrenz vorlagen.
- Um einen Ausgleich zu den mit Entitäten angereicherten Anteilen zu schaffen, wurden ca. 500 Titel und Sätze *ohne* Vorbedingung aus den PubMed Kurzzusammenfassungen randomisiert hinzugefügt.

Ob der auf diese Weise zusammengestellte *Annotationskorpus* alle gestellten Bedingungen erfüllte, konnte zu dem Zeitpunkt der Zusammenstellung nicht festgestellt werden. Dies wurde im Anschluss manuell überprüft. Die tatsächliche Zusammensetzung unterscheidet sich nach der manuellen Annotation dieses Textkörpers und wird im Kapitel Ergebnisse im Abschnitt *3.2. Betrachtung des Annotationskorpus* näher erläutert.

Im Gegensatz zu den anderen Textkörpern ist der *Annotationskorpus* statisch und wird nicht vor jedem neuen Suchdurchlauf neu gebildet. Die darin enthaltenen einzelnen Sätze und Titel gelten jeweils als ein Element. Für jedes Element wurden manuell alle Vorkommen von Enzymen, Krankheiten und Organismen annotiert. Bei Titeln und Sätzen, in denen eine Kookkurrenz von Enzym- und Krankheitsentitäten vorlag, wurde zusätzlich eine manuelle Klassifizierung der semantischen Entitätsrelation vorgenommen.

Basierend auf den Ergebnissen der manuellen Klassifizierung der semantischen Beziehung der Entitäten, wurden folgende Relationsklassen definiert: *causal interaction (kausale Interaktion)*, *ongoing research (Gegenstand der Erforschung)*, *diagnostic usage (diagnostische Nutzung)* und *therapeutic application (therapeutische Anwendung)*. Die Bedingungen für eine Einordnung einer Kookkurrenz werden im Abschnitt *2.4.1. Definitionen der Entitätsrelationen* eingehend erklärt. Der jeweiligen Klasse werden alle Titel und Sätze im *Annotationskorpus* zugeordnet, deren Kookkurrenz von Enzym- und Krankheitsentitäten, die dort beschriebene semantische Beziehung widerspiegeln. Dabei kann ein und dieselbe Kookkurrenz auch mehreren Klassen zugeordnet sein oder (bei entsprechender Zusammenhanglosigkeit) keiner der definierten Relationen.

2.1.2. Wörterbücher und Ausschlusslisten

Im Rahmen dieser Doktorarbeit wurden Wörterbücher verwendet, die als Grundlage für die Suche nach Krankheitsentitäten dienten und bei der Vorverarbeitung der zu klassifizierenden Daten eingesetzt wurden. Zur Optimierung des Krankheitswörterbuches und zur Qualitätsverbesserung der Ergebnisse der regelbasierten Identifizierung der Datenbankzugangsnummern wurden Ausschlusslisten angewendet. Daneben wurden für die Vorverarbeitung der zu klassifizierenden Objekte die Eignung von Stoppwortlisten getestet.

Wörterbuch mit Krankheiten

Die Suche nach Krankheiten im *Kookkurrenzkorpus* basiert auf einer wörterbuchbasierten Suche nach Entitäten. Bei einem wörterbuchbasierten Ansatz wird eine Sammlung von Begriffen, mit Namen und Synonymen für Entitäten, als Referenzwörterbuch zur Suche der entsprechenden Entität verwendet. Ein Treffer gilt als positive Bestätigung des Vorhandenseins der Entität. Ein Wörterbuch kann aber auch als Ausschlussliste verwendet werden, um z.B. Homonyme zu entfernen.

Die Sammlung der *Medical Subject Headings*, (MeSH) [26], als hierarchisch organisierter Thesaurus aus dem Bereich der Lebenswissenschaften und der Medizin, dient als Grundlage für die Erstellung des Krankheitswörterbuches. Bevor eine Entitätssuche durchgeführt wurde, ist immer eine tagesaktuelle Version der MeSH Begriffe bezogen und extrahiert worden. Die hierarchische Struktur von MeSH macht es möglich gezielt Krankheitsbegriffe aus der Sammlung zu extrahieren, denn nur diese werden für eine Suche nach Krankheitsentitäten benötigt. Begriffe mit mindestens einer Zeichenlänge von vier wurden für die Zusammenstellung des Krankheitswörterbuchs den entsprechenden MeSH Kategorien (Tabelle 2.1) entnommen.

Tabelle 2.1: Eine Auflistung aller Subkategorien und Namen der Kategorie C (Krankheiten) der MeSH Hierarchie. Die rechte Spalte gibt Auskunft darüber ob die Begriffe der Subkategorie für die Zusammenstellung des Wörterbuches berücksichtigt wurden. Die Begriffe der Subkategorien C24-26 wurden nur verwendet, wenn sie gleichzeitig der semantischen Kategorie T047 (Krankheit oder Syndrom) von MeSH zugeordnet sind. Das englische Original der Tabelle befindet sich im Anhang.

Name der Kategorie	Kategorie	verwendet
Bakterielle Infektionen und Mykosen	[C01]	ja
Virosen	[C02]	ja
Parasitäre Erkrankungen	[C03]	ja
Neoplasien	[C04]	ja
Muskuloskeletale Erkrankungen	[C05]	ja
Erkrankungen des Verdauungsapparates	[C06]	ja
Erkrankungen des Kiefers und der Mundhöhle	[C07]	ja
Atemwegserkrankungen	[C08]	ja
Erkrankungen des Hals-Nasen-Ohren-Bereichs	[C09]	ja
Erkrankungen des Nervensystems	[C10]	ja
Erkrankungen des Auges	[C11]	ja
Erkrankungen des männlichen Urogenitaltraktes	[C12]	ja
Erkrankungen des weiblichen Urogenitaltraktes, Schwangerschaftskomplikationen	[C13]	ja
Kardiovaskuläre Erkrankungen	[C14]	ja
Hämolytische und lymphatische Erkrankungen	[C15]	ja
Angeborene, erblich bedingte und neonatale Erkrankungen/Anomalien	[C16]	ja
Haut- und Bindegewebserkrankungen	[C17]	ja
Stoffwechselerkrankungen und ernährungsbedingte Störungen	[C18]	ja
Erkrankungen des endokrinen Systems	[C19]	ja
Erkrankungen des Immunsystems	[C20]	ja
Umweltbedingte Erkrankungen	[C21]	ja
Krankheiten aus der Veterinärmedizin	[C22]	nein
Pathologische Zustände, Anzeichen und Symptome	[C23]	nein
Berufskrankheiten	[C24]	(ja)
Substanz bedingte Syndrome und Suchterkrankungen	[C25]	(ja)
Wunden und Verletzungen	[C26]	(ja)

Bei Begriffen mit einer Zeichenlänge von drei und weniger handelt es sich um Akronyme, von denen angenommen wurde, dass sie nicht eindeutig für eine Krankheitsentität sind und so den Anteil falsch-positiver Ergebnisse stark erhöhen könnten. Ein Beispiel für ein solches Synonym ist „ARC“. Es ist bei MeSH aufgeführt als Akronym für „AIDS-Related Complex“. Diese Abkürzung wird in vielen Bereichen anderweitig verwendet. Einige Entitäten, für die das Akronym „ARC“ stehen kann,

kommen auch aus dem Bereich der Wissenschaft (z.B. Arthritis Research Campaign, Australian Research Council, American Red Cross) und können deswegen auch vermehrt in Kurzzusammenfassungen der PubMed Datenbank auftreten.

Die Begriffe in der MeSH Kollektion sind ausschließlich in ihrer amerikanisch-englischen (AE) Schreibweise aufgeführt. Die in PubMed enthaltenen wissenschaftlichen Referenzen können aber sowohl in amerikanischem als auch in britischem Englisch (BE) verfasst sein. Aus diesem Grund wurden Krankheitsbegriffe, die in beiden Sprachvarietäten unterschiedlich geschrieben werden, bei der Erzeugung des Wörterbuchs in der fehlenden Variante ergänzt. In Tabelle 2.2 sind zwei Beispiele für Unterschiede in der Schreibweise aufgeführt.

Tabelle 2.2: Beispiele für unterschiedliche Schreibvariationen für die selbe Krankheit in amerikanischem und britischem Englisch.

Britisches Englisch	Amerikanisches Englisch	Deutsch
anaemia	anemia	Anämie
tumour	tumor	Tumor

Das so erstellte Wörterbuch enthielt einige Begriffe, die zu falsch-positiven Zuordnungen führen könnten. Hierbei handelte es sich um mehrdeutige Wörter oder Begriffe. Das Wort „strain“ steht bei MeSH für eine Muskelzerrung. In der Systematik von Organismen bezeichnet es den Begriff „Stamm“. Während der manuellen Annotation (siehe2.1.1. Textkorpus) wurden solche mehrdeutigen Begriffe in einer Ausschlussliste gesammelt. Nach der Erstellung des Wörterbuchs, wurden abschließend solche Begriffe aus dem Krankheitswörterbuch gelöscht.

Das für die Suche nach Krankheitsentitäten generierte Wörterbuch umfasst 22.380 Begriffe zu 4.054 Krankheitsentitäten (Stand BRENDA Aktualisierung Januar 2011). Es enthält damit durchschnittlich zwischen fünf und sechs Synonyme für eine Krankheitsentität.

Wörterbuch mit Enzymen

Die Benennung von Bioentitäten in wissenschaftlichen Texten ist oft uneinheitlich. So werden auch Enzyme oft nicht mit dem durch die IUBMB empfohlenen Namen oder ihrer EC-Nummer benannt. Im Rahmen der halbjährlichen Aktualisierung von BRENDA wurden aus den ca. 100.000 (Stand BRENDA Aktualisierung Januar 2011) bekannten Enzymnamen und Synonymen ein Enzymwörterbuch generiert. Dieses Wörterbuch wird bei der bereits vor dieser Arbeit etablierten Text Mining basierten

Suche nach Enzymentitäten [46] in PubMed Kurzzusammenfassungen generiert (Abbildung 2.1) und wurde daraus übernommen und verwendet. In BRENDA sind für eine Enzymentität durchschnittlich 15 Synonyme bekannt [43]. Das Enzymwörterbuch enthält 58.646 Begriffe zu 4.618 Enzymentitäten (Stand BRENDA Aktualisierung Januar 2011). Bei allen Prozessierungsschritten, die Namen und Synonyme von Enzymen benötigten, wurde die jeweils aktuelle Version dieses Wörterbuchs verwendet.

Ausschlusslisten

Wie bei der Zusammenstellung des Krankheitswörterbuchs, wird auch bei der Suche nach Zugangsnummern von wissenschaftlichen Datenbanken eine Ausschlussliste verwendet (siehe *2.3. Identifizierung von Entitäten* auf Seite 25). Die Suche selbst arbeitet regelbasiert und nicht unter Verwendung eines Wörterbuchs. Nach der Suche wurden alle Ergebnisse mit einer manuell erstellten Ausschlussliste gefiltert, um falsche Zuordnung zu vermeiden. Diese Ausschlussliste umfasst unter anderem Begriffe, die den Regeln einer Formatierung von Zugangsnummern entsprechen aber bekannte Synonyme für Enzyme oder Abkürzungen für chemische Substanzen sind.

Stoppwortlisten

Um bei der Vorverarbeitung der Titel und Sätze bei der Klassifizierung die Gesamtanzahl der zu verarbeitenden Wörter zu reduzieren, wurde die Filterung mit zwei Stoppwortlisten getestet. Stoppwörter sind allgemeine Begriffe, die keine themenspezifischen Inhalte vermitteln, wie Artikel (the, a), Konjunktionen (and, or) oder Präpositionen (by, at).

Ausgehend von einer statischen Stoppwortliste[4] des SMART Retrieval Systems [47], die Stoppwörter der englischen Sprache enthält, wurde diese nach einer Analyse der Wortfrequenz in den PubMed Kurzzusammenfassen (*3.1.1. Wörter in PubMed Titeln und Kurzzusammenfassungen* auf Seite 47) angeglichen. Es wurden zwei Versionen einer angepassten Stoppwortliste auf ihren Nutzen bei der Vorverarbeitung der Klassifizierungsdaten getestet. Eine Tabelle mit den Wörtern der Stoppwortlisten sind im Anhang zu dieser Arbeit aufgeführt.

[4] ftp://ftp.cs.cornell.edu/pub/smart/english.stop

Wörterbuch des *Kookkurrenzkorpus*

Der *Kookkurrenzkorpus* liegt nicht in einer für das menschliche Auge lesbare Form vor, sondern als Hashwertrepräsentation (*2.2.1. Hashwerte statt natürliche Sprache*). Für die Identifizierung von Krankheitsentitäten und die Suche nach Kookkurrenzen mit Enzymentitäten ist die Ursprungsinformation, also das natürliche Wort als Textrepräsentanz, nicht mehr von Bedeutung. Dennoch wird in Fällen, in denen das Ergebnis der Kookkurrenzsuche eine Grundlage für eine weitere manuelle Auswertung bildet, wie zum Beispiel beim Aufbau eines manuell annotierten Test- und Trainingskorpus, wieder eine für das menschliche Auge lesbare Repräsentation im Reintextformat benötigt. Zu diesem Zweck ist aus dem Textkorpus der PubMed Titel und Kurzzusammenfassungen ein Wörterbuch aller vorkommenden Wörter extrahiert worden. Diesen Wörtern wurden die jeweiligen Hashwerte, die für sie aus einer Hashfunktion erzeugt wurden, zugeordnet. So kann jedes Wort des *Kookkurrenz-* und *Klassifikationskorpus* von seiner Hashwertform in seine alphanumerische Textform überführt werden.

2.2. Textaufbereitung und Datenstrukturen

Um Techniken und Methoden aus dem Bereich Text Mining anzuwenden ist eine Vorverarbeitung elektronischer Dokumente unumgänglich [48]. Welche Art der Vorverarbeitung gewählt wird, kann durch die jeweilige angewandte Methode, den Umfang der Ausgangsdaten und auch die Sprache beeinflusst werden [49]. Bei der Vorverarbeitung werden elektronische Dokumente in Repräsentationsformen umgewandelt, die weitere Verarbeitungsschritte ermöglichen beziehungsweise verkürzen.

In den folgenden Abschnitten wird beschrieben, in welcher Form die jeweils verwendeten Textkörper ursprünglich vorlagen, welche Art der Verarbeitung durchgeführt wurde und in welche Datenstrukturen sie durch diese Verarbeitung überführt wurden. Des Weiteren wird erklärt, in welcher Form die Ergebnisse der Analysen gespeichert wurden.

2.2.1. Hashwerte statt natürliche Sprache

Der *Kookkurrenzkorpus* wird durch das bereits vor dieser Arbeit etablierter Text Mining Verfahren aus den PubMed Kurzzusammenfassungen im Rahmen der halbjährlichen Aktualisierung von BRENDA gebildet [46] und wurde als Material aus dieser Quelle übernommen und verwendet (Abbildung 2.1). Im Zuge seiner Erstellung wird der

Kookkurrenzkorpus nach Sätzen getrennt und diese wiederum nach Wörtern. Die Wörter werden als sogenannter Hashwert (Streuwert) in einer Datenbank abgelegt. Der Hashwert ist ein Ergebnis einer Hashfunktion. Unter einer Hashfunktion versteht man einen Algorithmus, der einem Wert einer Eingabemenge einen Wert einer Zielmenge zuordnet, den Hashwert. Die Eingabemenge bilden alle vorkommenden Wörter des *Kookkurrenzkorpus*, die aus alphanumerischen Zeichenketten bestehen. Die Zielmenge bilden die erzeugten Hashwerte. Die Hashwerte sind rein numerisch aufgebaut. Es werden die Information über die Ursprungskurzusammenfassung, die Position des Wortes und der erzeugte Hashwert gespeichert. Die alphanumerische Eingabemenge, also das Wort an sich, wird nicht gespeichert.

Entitätssuche

Der so vorverarbeitete *Kookkurrenzkorpus* war das Ausgangsmaterial für die im Rahmen dieser Arbeit durchgeführte Suche nach Krankheitsentitäten. Das Krankheitswörterbuch (*2.1.2. Wörterbücher und Ausschlusslisten*) wurde ebenfalls durch eine Hashfunktion in Hashwerte überführt. Das Auftreten einer Entität im Text wurde durch den Vergleich der Hashwerte erfasst und die Information über die Ursprungsreferenz und die Position gespeichert.

Kookkurrenzsuche

Durch die Ergebnisse des Text Mining Verfahrens von BRENDA waren die Positionen der Enzymentitäten bekannt (Abbildung 2.1). Es erfolgte ein Abgleich der Positionsinformationen. Die ermittelten Positionen der Krankheitsentitäten und die Positionen der Enzymentitäten wurden verglichen und die Positionen von Kookkurrenzen in einem Satz oder einem Titel gespeichert. Ebenso wie bei der Entitätssuche, ist für die Kookkurrenzsuche die Ursprungsinformation, also das Wort als Textrepräsentanz, nicht mehr von Bedeutung.

2.2.2. Extraktion des Textanteils

Die Suche nach Zugangsnummern biologischer Datenbanken fand in dem extrahierten *Volltextkorpus* (*2.1.1. Textkorpus*) statt. Um nach den Zugangsnummern mit der vorgesehenen Methode (*2.3.1. Regelbasierter Ansatz*) suchen zu können, wurde der Textanteil extrahiert. Der Textanteil entspricht allen Zeichenketten, ohne Graphiken und

ohne solche Tabellen, die nur als Graphiken gespeichert vorliegen. Bei Tabellen, die extrahierbare Textanteile enthalten geht bei diesem Vorgang allerdings die Information der Formatierung verloren.

Bereits vorher bekannte Informationen über die Referenz, z.B der entsprechenden PubMed Identifikationsnummer, dem Namen der Zeitschrift und der EC-Nummer, über die der Artikel relevante Informationen enthält, wurden in einer Datenbank abgelegt. Der Textanteil wurde anschließend auf das Vorhandensein von Zugangsnummern untersucht. Die gefundenen Zugangsnummern wurden mit einem Verweis auf die Ursprungsreferenz gespeichert.

2.2.3. Aufbereitung und Termgewichtung

Als Ausgangsdaten für die Klassifizierung diente der *Klassifikationskorpus*. Er bildet eine Teilmenge des *Kookkurrenzkorpus* und liegt ebenso wie dieser, als Hashwertrepräsentation des alphanumerischen Ursprungstextes vor (*2.2.1. Hashwerte statt natürliche Sprache*). Um als Eingabeinstanz von der Support Vector Machine (SVM) (*2.4.2. Relationsklassifizierung mit Support Vector Machines*) verarbeitet werden zu können, musste er weiter aufbereitet werden. Eine Einordnung dieser Schritte in den Gesamtverlauf der für die Klassifizierung durchgeführten Verarbeitung ist schematisch in Abbildung 2.6 auf Seite 32 dargestellt.

Der erste Schritt der Aufbereitung wurde in zwei verschiedenen Varianten getestet. Die eine Variante bestand aus einer Entfernung aller Namen der gefundenen Enzym- und Krankheitsentitäten in den zu klassifizierenden Sätzen und Titeln (*Löschung*). Bei der zweiten Variante wurden die Namen der Entitäten nicht entfernt sondern ersetzt. Alle Enzymentitäten wurden mit dem generischen Enzymnamen „enyzme-xyz“, alle Krankheitsentitäten mit dem generischen Krankheitsnamen „disease-xyz“ benannt (*Austausch*). Dieses Vorgehen sollte die Gewichtung der jeweiligen Namen bei der späteren Berechnung der Termgewichte abschwächen.

In einem weiteren Verarbeitungsschritt wurden die Termgewichte aller Wörter des *Klassifikationskorpus* berechnet. Durch eine Berechnung von Termgewichten kann festgelegt werden, wie relevant ein einzelnes Wort und sein Informationsgehalt im Bezug auf den gesamten Textkörper und das jeweilige Element (Satz, Titel, Dokument) eines Textkörpers ist. Jeder *Satz* oder *Titel* des *Kookkurrenzkorpus* gilt als Informationseinheit und jedes *Wort* als Untereinheit der Informationseinheit. Die Informationseinheiten wurden zu einem Vektor in einem multidimensionalen Raum umgerechnet und damit zu einer abstrakten Repräsentation im Vektorraum. Diese Repräsentation wurde nach dem *Vector Space Model* [50] erstellt. Die Anzahl der

Koordinaten eines Vektors richtete sich nach der Anzahl der *Wörter* in der Informationseinheit *Satz/Titel*. Für jedes *Wort* wird ein Termgewicht errechnet. Das Termgewicht ist ein Indikator für den Informationswert des Wortes. Alle Termgewichte zusammen bilden den jeweiligen Vektor der Informationseinheit, und alle Vektoren spannen zusammen den multidimensionalen Vektorraum auf.

Als Methode zur Bestimmung der Termgewichte wurde die Berechnung der Termfrequenz tf_{ij} unter Berücksichtigung der inversen Dokumentfrequenz idf_i angewandt [51], im Weiteren hier kurz *tf-idf* genannt. Die Termfrequenz tf_{ij} ist die Anzahl der Vorkommen eines Worts t_i in der Informationseinheit d_j. Die inverse Dokumentfrequenz idf_i wird berechnet als

$$idf_i = \ln\left(\frac{D}{1+d_{it}}\right) \quad (2.1)$$

wobei D die Anzahl aller Informationseinheiten ist und d_{it} die Anzahl der Informationseinheiten, die Wort t_i enthalten, beschreibt. Die *tf-idf* ist das Produkt von tf_{ij} und idf_i.

$$tf * idf_{ij} = tf_{ij} * idf_i \quad (2.2)$$

Im Anschluss wurden alle Vektoren zu Einheitsvektoren umgerechnet, indem alle Koordinaten durch die Länge des Vektors geteilt wurden. So verliert die unterschiedliche Anzahl von Wörtern in Titeln und Sätzen ihren Einfluss auf die Gewichtung, und deshalb entkoppelt die Normierung der Vektoren die Gewichtung der Worte von den unterschiedlichen Längen der Informationseinheiten [52].

2.3. Identifizierung von Entitäten

In diesem Abschnitt werden die Methoden zur Bestimmung von gemeinsam auftretenden Enzymen und Krankheiten, sowie der Identifizierung von Zugangsnummern biologischer Datenbanken beschrieben. Die Suche nach Zugangsnummern wissenschaftlicher Datenbanken erfolgte über die in Abschnitt 2.3.1. erläuterte regelbasierte Methode. Die Identifizierung von Krankheiten und das Auffinden einer Kookkurrenz mit einem Enzym beruht auf einem wörterbuchbasierten Ansatz (Abschnitt 2.3.2.).

2.3.1. Regelbasierter Ansatz

Die Einträge in unterschiedlichen wissenschaftlichen Datenbanken besitzen ein uneinheitliches Format und können die verschiedensten Datenfelder enthalten. Um einen bestimmten Eintrag zu referenzieren, haben Datenbanken einen eindeutigen Identifikator. Dieser Identifikator ist meist eine alphanumerische Zeichenfolge und wird auch *Zugangsnummer* (eng. accession number) genannt.

In Texten mit naturwissenschaftlichen Hintergrund wird zur eindeutigen Bestimmung einer beschriebenen Entität, z.B. eines Proteins, häufig begleitend die entsprechende Zugangsnummer einer Datenbank angegeben. Die computergestützte Erfassung der im Text enthaltenen Zugangsnummern kann helfen, Teilaspekte der Texte automatisch zu erschließen. Die Identifizierung einer Zugangsnummer, deren Bedeutungsaufschlüsselung in der Regel nur mit der Kenntnis des Datenbankeintrags möglich ist, wäre über ein zu erstellendes Wörterbuch nur mit erheblichem Aufwand zu realisieren. Deswegen wurde ein regelbasierter Ansatz verfolgt, um Zugangsnummern wissenschaftlicher Datenbanken in den wissenschaftlichen Artikeln des *Volltextkorpus* (siehe Textkorpus auf Seite 14) zu identifizieren.

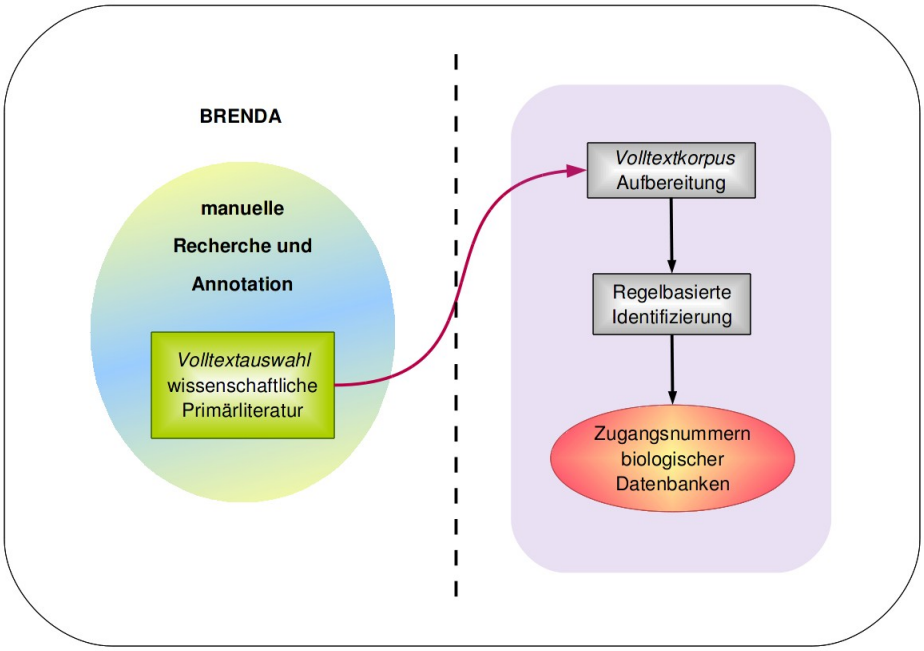
In Tabelle 2.3 sind die Datenbanken aufgeführt, nach deren Zugangsnummern in Texten gesucht wurde. Die *International Nucleotide Sequence Database Collaboration* (INSDC) [53] umfasst eine Kollaborationsgemeinschaft von verschiedenen Instituten bestehend aus den Sequenzdatenbanken: *DNA Databank of Japan* (DDBJ) [54], des *European Molecular Biology Laboratory's European Bioinformatics Institute* (EMBL-EBI) [55] und der *GenBank* [39] des NCBI, die auch auf der Ebene der Formatvorgaben und der Vergabe von Zugangsnummern kooperiert. Die Formatierungsvorgaben der INSDC und der anderen Datenbanken (Tabelle 2.3), nach deren Zugangsnummern gesucht wurde, wurden aus den Dokumentationen und Hilfeseiten der jeweiligen Datenbanken extrahiert und in ein Regelsystem für die alphanumerischen Struktur überführt.

Tabelle 2.3: Die lebenswissenschaftlichen Datenbanken deren Zugangsnummern im Volltextkorpus gesucht wurden. Neben den Namen der Datenbanken ist deren Inhaltsschwerpunkt sowie die Adresse der Internetpräsenz aufgeführt.

Name der Datenbank	Inhaltsschwerpunkt	Internetpräsenz [Referenz]
DNA Data Bank of Japan (DDBJ)	Sequenzdatenbank (Nukleotide)	http://www.ddbj.nig.ac.jp [54]
EMBL Nucleotide Sequence Database	Sequenzdatenbank (Nukleotide)	http://www.ebi.ac.uk/embl [55]
GenBank	Sequenzdatenbank (Nukleotide)	http://www.ncbi.nlm.nih.gov/genbank [39]
NCBI RefSeq	Sequenzdatenbank (DNA, mRNA, Proteine)	http://www.ncbi.nlm.nih.gov/RefSeq [56]
Protein Data Bank (PDB)	3D-Strukturdaten von Proteinen und Nukleinsäuren	http://www.pdb.org [57]
PROSITE	Datenbank für Protein-Familien und Domänen	http://www.expasy.org/prosite [42]
SwissProt/UniProt/TrEMBL	Sequenzdatenbank (Proteine)	http://www.uniprot.org [40]

Durch die Kenntnis über die Quelldatenbank und anhand des Formats der Zugangsnummer ist es möglich die Art des Eintrags zu bestimmen, d.h. ob es sich um die Zugangsnummer zu einer Protein- oder Nucleotidsequenz oder einer Proteinstruktur handelt. Entsprechend diesen Regeln wurden reguläre Ausdrücke formuliert. Ein regulärer Ausdruck ist eine definierte syntaktische Regel, die eine Zeichenkette oder eine Menge von Zeichenketten definiert und so als Filterkriterium oder Schablone dienen kann. Eine ausführliche Auflistung aller definierten Regeln und regulärer Ausdrücke für die Datenbanken befindet sich im Anhang in Tabelle Anhang 5 und Tabelle Anhang 6 auf den Seiten 99 und 100.

Abbildung 2.4: Eine schematische Darstellung der Schritte der regelbasierten Suche nach Zugangsnummern biologischer Datenbanken. Im rechten Bildbereich sind die in dieser Arbeit etablierten Verarbeitungsschritte (lila unterlegt) und im linken Bildbereich (blau-gelb unterlegt) die genutzten Ressourcen des BRENDA Informationssystems dargestellt.



Es wurde nach der im regulären Ausdruck definierten Menge von Zeichenketten gesucht (Abbildung 2.4). Daneben wurde auch nach Indikatortermini gesucht. Ein Indikatorterm ist ein Wort oder eine Wortfolge, die die korrekte Einordnung als Zugangsnummer sowie die Zuordnung zur entsprechenden Datenbank bestätigen soll. Die Information über die gefundene Zugangsnummer und die Distanz zu einem Indikatorterm werden in einer Ergebnisdatenbank abgelegt.

Trotz der eindeutigen Definitionen der regulären Ausdrücke können auch Zeichenfolgen, die keine Zugangsnummern sind, als solche auftreten. Es handelt sich dann um Bezeichnungen, die ihrer Zeichenstruktur nach eine gültige Zugangsnummer wären. Es kann sich dabei z.B. um ein Synonym für ein Enzym handeln. Die Zeichenfolge „BGLU46“ würde dem gültigen Format für eine DDBJ Zugangsnummer entsprechen. Es wird aber auch als ein Synonym für Coniferin beta-Glukosidase (EC 3.2.1.126) genutzt, einem Enzym in der Pflanzenzellwand, das an der Biosynthese von Lignin beteiligt ist. In DDBJ findet man hierzu keinen Eintrag unter dieser Zeichenfolge als Zugangsnummer. Die Zeichenfolge ist also nur scheinbar eine DDBJ

Zugangsnummer. Deshalb wurden nach der Suche alle Ergebnisse durch eine manuell erstellte Ausschlussliste gefiltert. In der Ausschlussliste befinden sich Begriffe, die den Regeln einer Formatierung von Zugangsnummern entsprechen, aber z.B. allgemein verwendete Synonyme für Enzyme oder Abkürzungen für chemische Substanzen sind sowie Zugangsnummern, die auf Listen mit gelöschten/ungültigen Zugangsnummern der entsprechenden Datenbanken enthalten sind.

2.3.2. Wörterbuchbasierter Ansatz

Im Rahmen der halbjährlichen Aktualisierung der manuell annotierten Daten der BRENDA Datenbank werden zusätzlich seit 2006 zwei Datenbanken, AMENDA und FRENDA, durch ein automatische Auswertung der PubMed Kurzzusammenfassungen als komplementäre Informationsquellen erzeugt [45]. Es erfolgt bei diesem Verfahren eine Suche nach Enzym- und Organismenentitäten (Abbildung 2.1). Ergänzend dazu findet eine Suche nach Aussagen zu der subzellulären Lokalisierung und den die Enzyme enthaltenden Gewebearten in den jeweiligen Organismen statt [45]. Beide Suchansätze sind wörterbuchbasiert.

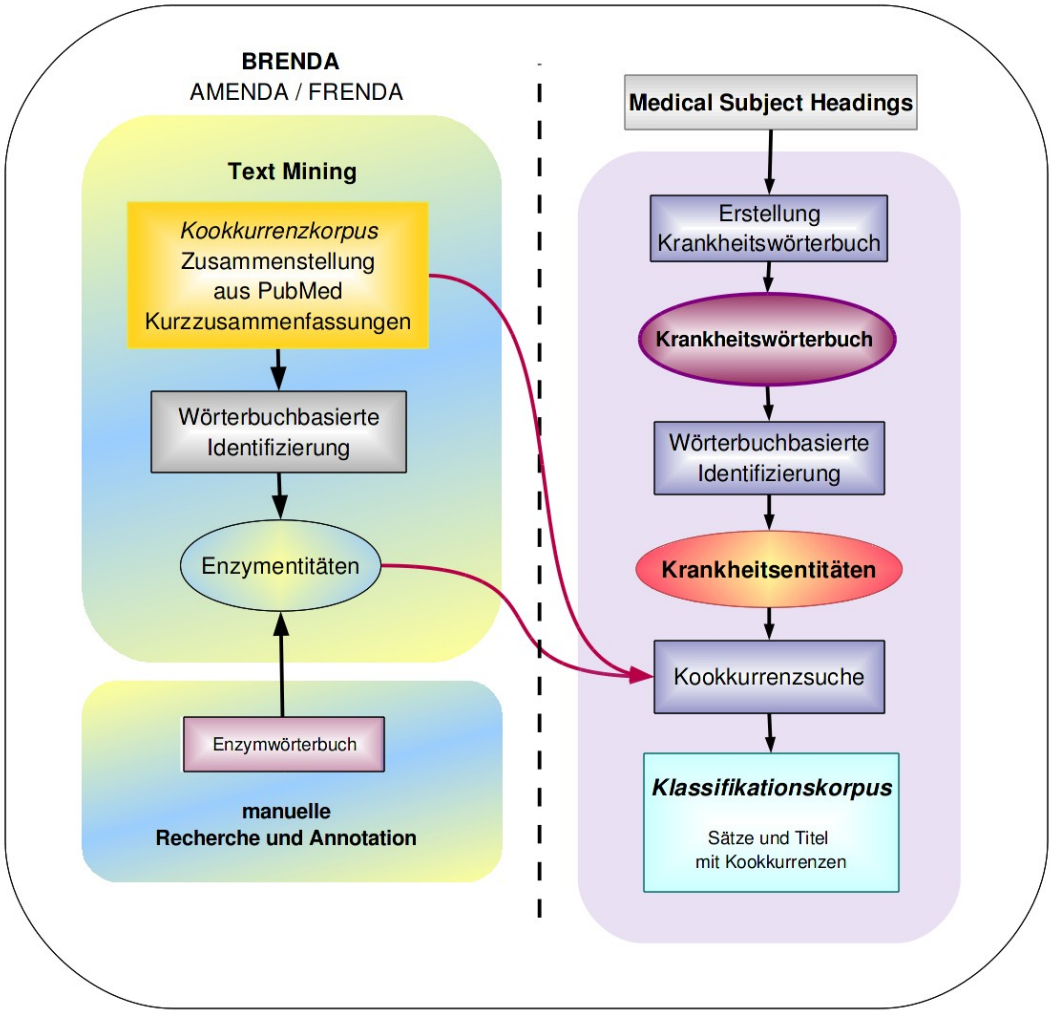
Eine Methode, die auf der Zuordnung von Phrasen in den PubMed Kurzzusammenfassungen zu Konzepten aus dem Unified Medical Language System (UMLS) [58] unter Verwendung des Programms MetaMap [59] basiert, diente bis vor Beginn dieser Arbeit der automatischen Suche nach Beziehungen zwischen Krankheiten und Enzymen [60,61]. Die Identifizierung von Krankheitsentitäten wurde während dieser Arbeit mit einem anderen Ansatz neu implementiert [43]. Dieser neue Ansatz orientiert sich an der Methode der automatischen Auswertung, wie sie bereits Bestandteil der BRENDA Aktualisierungsroutine war (Abbildung 2.1). Eine Übersicht der Schritte der nun wörterbuchbasierten Identifizierung von Krankheitsentitäten ist schematisch in Abbildung 2.5 dargestellt.

Identifizierung von Krankheitsentitäten

Für die Suche nach Krankheitsentitäten (Abbildung 2.5) wurde die bereits in BRENDA etablierte Methode der wörterbuchbasierten Suche nach Entitäten auf Krankheitsentitäten ausgeweitet. Dazu wurde ein eigenes Krankheitswörterbuch zusammengestellt (*2.1.2. Wörterbücher und Ausschlusslisten*), anhand dessen Namen von Entitäten in dem *Kookkurrenzkorpus* gesucht wurden. Die Namen können aus ein bis sechs Wörtern bestehen, das heißt, es werden maximal Hexagramme von Wörtern

berücksichtigt. Die Wörter des *Kookkurrenzkorpus* und des Wörterbuchs sind jedoch nicht als Text, sondern als Hashwert einer Hashfunktion gespeichert (*2.2.1. Hashwerte statt natürliche Sprache*).

Abbildung 2.5: Eine schematische Darstellung der Schritte, die zu der Identifizierung der Krankheitsentitäten führen und die anschließenden Kookkurrenzsuche ermöglichen. Im rechten Bildbereich sind die in dieser Arbeit etablierten Verarbeitungsschritte (lila unterlegt), im linken Bildbereich (blau-gelb unterlegt) sind die genutzten Ressourcen des BRENDA Informationssystems dargestellt. Zur Bildung des Krankheitswörterbuches wurde auf den MeSH Thesaurus zurückgegriffen.



Bestimmung von auftretenden Kookkurrenzen

Eine Kookkurrenz ist in der Linguistik dadurch definiert, dass zwei namentlich genannte Entitäten innerhalb eines definierten sprachlichen Rahmens (z.B. Satz, Absatz oder Dokument) mindestens einmal gemeinsam auftreten [62]. Eine Kookkurrenz wird als ein Indiz für eine grammatische oder semantische Abhängigkeit zwischen den zwei Entitäten gedeutet.

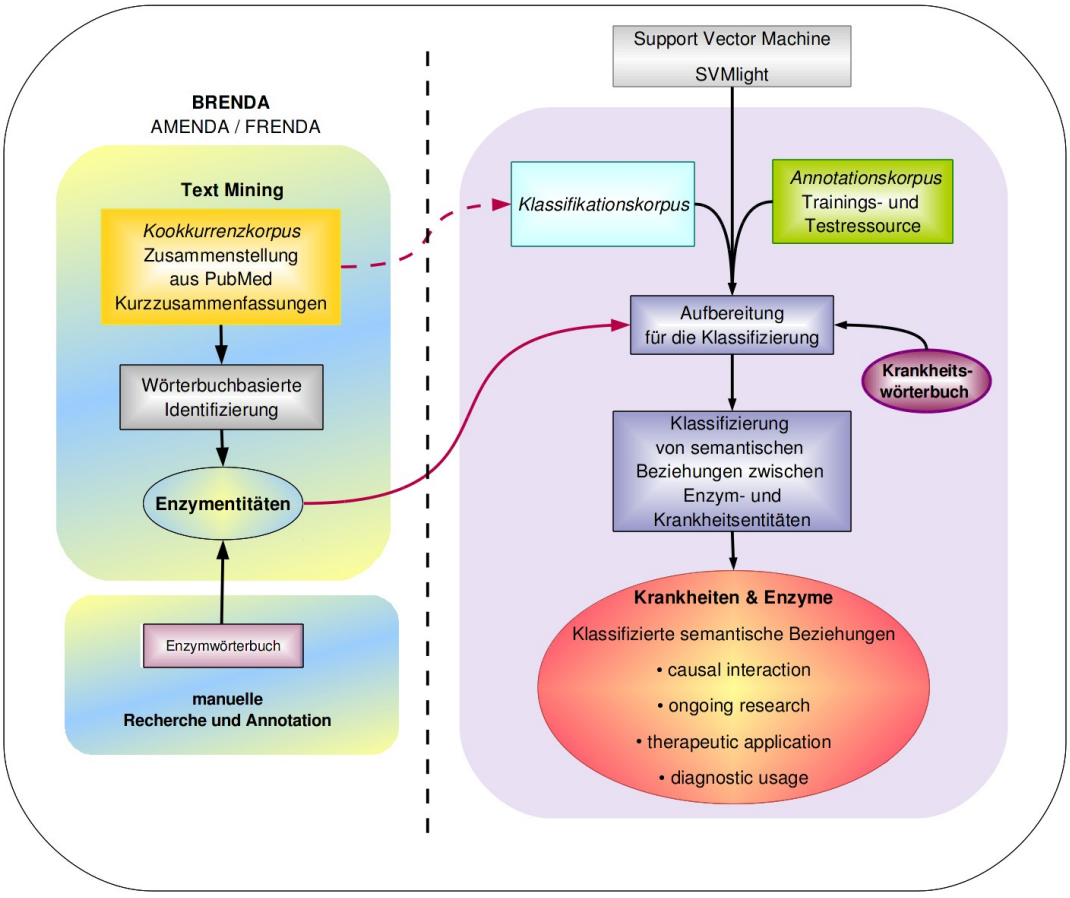
In dieser Arbeit sollten neben dem Auftreten der Krankheitsentitäten auch das Auftreten von Kookkurrenzen dieser mit einer Enzymentität bestimmt werden. Durch die Ergebnisse des Text Mining Verfahrens von BRENDA waren die Positionen der Enzymentitäten bekannt. Diese bildeten einen Teil der Quellinformation für die Bestimmung der Kookkurrenzen mit Krankheitsentitäten. Nach der durchgeführten Identifizierung von Krankheitsentitäten (Abbildung 2.5) wurden die Ergebnisse mit den Positionsinformationen der Enzymentitäten abgeglichen und das gemeinsame Auftreten in einem Satz oder einem Titel in dem *Kookkurrenzkorpus* festgehalten.

2.4. Klassifizierung der semantischen Entitätsbeziehungen

Allein das Wissen um eine örtliche Nähe zweier Entitäten sagt nichts über ihre Beziehung zueinander aus. Die gemeinsame Nennung von Enzymen und Krankheiten in einem Satz könnte auf eine engere semantische Verbundenheit hindeuten, ist aber allein daraus nicht näher differenzierbar. Sofern man kookkurrierende Entitäten in Texten findet, müssen sie also weiter untersucht werden, um einen semantischem Zusammenhang zu bestimmen.

Um eine Aussage über die Beziehung zweier Entitäten machen zu können, wurden semantische Relationsklassen (*2.4.1. Definitionen der Entitätsrelationen*) definiert, in die mit Hilfe einer Methode aus dem Bereich des maschinellen Lernens (*2.4.2. Relationsklassifizierung mit Support Vector Machines*) die gefundenen kookkurrierenden Entitäten eingeordnet wurden. Eine Übersicht der Schritte, die auf dem Weg zu einer der Klassifizierung der semantischen Beziehungen durchlaufen werden, ist schematisch in Abbildung 2.6 dargestellt.

Abbildung 2.6: Eine schematische Darstellung der Schritte, die für eine Klassifizierung der semantischen Beziehungen von kookkurrierenden Krankheits- und Enzymentitäten durchlaufen werden. Im Abschnitt 2.4.1. „Definitionen der Entitätsrelationen“ ab Seite 33 sind die definierten semantischen Beziehungen eingehend erläutert. Vorausgehend ist der Schritt der Aufbereitung bereits im Abschnitt 2.2.3. „Aufbereitung und Termgewichtung “ behandelt worden. Im linken Bildbereich (blau-gelb unterlegt) sind die genutzten Ressourcen des BRENDA Informationssystems dargestellt, im rechten Bildbereich die in dieser Arbeit etablierten Verarbeitungsschritte (lila unterlegt). Die gestrichelte rote Linie soll andeuteten, dass der Klassifikationskorpus bereits aus dem Verarbeitungsschritten der Kookkurrenzsuche hervorgegangen ist (Abbildung 2.5). Die zur Klassifizierung verwendete Methode des maschinellen Lernens basiert auf der Theorie der Support Vector Machine (SVM). Der theoretische Hintergrund ist in Abschnitt 2.4.2. „Relationsklassifizierung mit Support Vector Machines“ näher erläutert. Das Programm SVMlight [68] wurde für die SVM basierte Verarbeitung eingebunden.



2.4.1. Definitionen der Entitätsrelationen

Um eine Klassifizierung vorzunehmen, wurden vier semantische Relationen definiert. Diese sind im Einzelnen *causal interaction* (kausale Interaktion), *ongoing research* (Gegenstand der Erforschung), *diagnostic usage* (diagnostische Nutzung) und *therapeutic application* (therapeutische Anwendung):

- ***Causal interaction (kausale Interaktion)***: Die kritische Rolle von Enzymen in katalysierten metabolischen Reaktionen impliziert, dass eine Fehlfunktion häufig pathologische Zustände von Organismen verursachen kann. Die Fehlfunktionen können, z.B. durch eine Mutationen im kodierenden Gen und einer dadurch veränderten Aminosäuresequenz, direkt verursacht werden. Als weitere Ursachen für eine Beeinträchtigung der katalytischen Funktion des Enzyms kommen die Präsenz eines Inhibitors (z.B. Nebenwirkungen von Medikamenten) oder das Fehlen von benötigten Kofaktoren in Betracht. Der Klasse *causal interaction* werden alle Titel und Sätze zugeordnet, deren Kookkurrenz von Enzym- und Krankheitsentitäten gleichzeitig diese semantische Verknüpfung beinhaltet.

Beispielsatz:

> *„Chronic granulomatous disease (CGD) results from mutations of phagocyte NADPH oxidase.“ [63]*

- ***Ongoing research (Gegenstand der Erforschung)***: Während die wissenschaftliche Forschung voranschreitet, werden mehr und mehr Zusammenhänge von Enzymen und Krankheiten, sei es als Auslöser, mögliches Therapeutikum oder diagnostisches Hilfsmittel, bekannt. In den in *PubMed* enthaltenen Referenzen finden sich auch Publikationen, in denen eine Wechselwirkung zwischen der enzymatischen Funktion oder Fehlfunktion und der Entwicklung einer Krankheit postuliert wird. In diesen Fällen werden die signifikanten Zusammenhänge zwar angenommen, sie sind jedoch noch nicht bewiesen und bedürfen der weiteren Erforschung.

Beispielsatz:

> *„The prognostic significance of epidermal growth factor receptor (EGFR) expression in lung cancer and, more importantly, its ability to predict response to anti-EGFR therapies, are currently subjects of active research.“ [64]*

- ***Diagnostic usage (diagnostische Nutzung):*** Im klinischen Labor werden eine Vielzahl von diagnostischen Parametern bestimmt. Diese können zur Abklärung eines Verdachts auf eine bestehende Erkrankung, zur Feststellung des Krankheitsstatus oder zur begleitenden Kontrolle eines Krankheitsverlaufs beitragen. Dabei gibt es viele diagnostische Verfahren, die die allgemeine Funktion oder spezifische Aktivität des Enzyms bestimmen. Die Präsenz von Enzymen kann z.B. auf Tumore hindeuten. Des Weiteren können veränderte Aktivitäten Indikator einer Organfehlfunktion sein.

Beispielsatz:

> *„Prostate-specific antigen (PSA) is the most clinically useful tumour marker available today for the diagnosis and management of prostate cancer.“ [65]*

- ***Therapeutic application (therapeutische Anwendung):*** Enzyme können eine wichtige Rolle in der Entwicklung und dem Fortschreiten einer Krankheit spielen, daher sind sie auch aus therapeutischer Sicht relevant. Einerseits können Enzyme Ziel eines Medikamentenwirkstoffs sein, andererseits können Enzyme selbst pharmakologisches Mittel zur Behandlung einer Krankheit sein.

Beispielsatz:

> *„Indinavir sulfate is a human immunodeficiency virus type 1 (HIV-1) protease inhibitor indicated for treatment of HIV infection and AIDS in adults.“ [66]*

Ziel der Klassifizierung ist es alle Titel und Sätze, deren Kookkurrenz von Enzym- und Krankheitsentitäten die beschriebene semantische Verknüpfung beinhalten, zu erfassen. Dabei kann ein und dieselbe Kookkurrenz auch mehreren Klassen entsprechen. Der folgende Beispielsatz kann den Klassen *ongoing research* sowie *therapeutic application* zugeordnet werden kann:

> *„Although the serine protease, tissue plasminogen activator (tPA), is approved by the US Food and Drug Administration for therapy to combat focal cerebral infarction, the basic concept of thrombolytic tPA therapy for stroke was challenged by recent studies that used genetically manipulated tPA-deficient (tPA-/-) mice, which suggested that tPA mediates ischemic neuronal damage.“ [67]*

2.4.2. Relationsklassifizierung mit Support Vector Machines

Um einen semantischen Zusammenhang automatisch erkennen zu können, kann man sich verschiedener Klassifikationsverfahren bedienen. Die Support Vector Machine (SVM) ist ein mathematisches Verfahren des statistischen Lernens zur automatischen Einteilung von unbekannten Objekten in Klassen durch Verwendung von Objekten bekannter Klassifizierung [19]. In dieser Arbeit wurde für die SVM gestützte Klassifizierung das Programm *SVMlight* [68] eingebunden.

Funktionsweise

Die SVM ist ein linearer Klassifikator. Eine SVM trennt Objekte in zwei Klassen, welche repräsentiert als Vektoren, einen Vektorraum aufspannen. Die Trennung erfolgt durch eine Hyperebene. Eine Hyperebene hat genau eine Dimension weniger als der Vektorraum und somit auch eine Dimension weniger als der Vektor mit der höchsten Dimension, der in diesem Raum dargestellt wird. Die Hyperebene wird durch die Vektoren (Stützvektoren) an ihrem Rand definiert. Die Bedingung dabei ist, dass der Algorithmus der SVM eine Hyperebene findet, deren Stützvektoren der beiden Klassen einen möglichst großen Abstand haben. Dieses Verfahren wird als Training der SVM bezeichnet und geschieht mit Objekten deren Klassenzugehörigkeit bereits vorher bekannt ist. Bei Objekten unbekannter Klassifizierung, die auf die gleiche Weise als Vektoren repräsentiert sind, soll die gefundene Hyperebene eine Klassifizierung ermöglichen. Die relative Position der Vektoren zur Hyperebene definiert, ob das Objekt einer Klasse angehört oder nicht.

Die SVM als Klassifikator von Texten

Bei der Verarbeitung von Texten und der dabei notwendigen Repräsentation (*2.2.3. Aufbereitung und Termgewichtung*) ist das Resultat jedoch ein hochdimensionaler Vektor, der nicht linear separierbar ist, weil die Vektoren der unterschiedlichen Klassen überlappen. Eine SVM ist dennoch ein geeigneter Klassifikator, wenn man die Dimension des Vektorraums soweit erhöht, bis eine lineare Trennung möglich ist. Berechnungen in einem hochdimensionalen Raum könnten, wegen des Rechenaufwands, ineffizient sein [69]. Bei einer SVM verwendet man deswegen einen Funktionskern (Kernel). Es gibt unterschiedlich definierte Kernel, wie z.B. lineare, polynomielle oder Kernel mit einer radialen Basisfunktion, die sich über unterschiedliche Parameter anpassen lassen. Neben der Verwendung eines Kernels kann der SVM erlaubt werden, durch eine Kosten-Nutzen-Anpassung bei der Wahl der

Hyperebene eine gewisse Anzahl von Fehlklassifizierungen zuzulassen. Zusammengefasst bedeutet es, dass der Algorithmus der SVM eine Möglichkeit mit einer minimalen Anzahl von Fehlklassifizierungen sucht, die Repräsentanten zweier Klassen optimal zu trennen. Dabei gilt, dass diese Trennung auch für unbekannte Objekte gültig bleiben sollte [69].

Trainieren der SVM

Durch das Trainieren der SVM entsteht ein Klassifikationsmodell, welches die Information über die Hyperebene und die Trennung der Klassen enthält. Die Ausgestaltung des Klassifikationsmodells, z.B. die zulässige Anzahl an Fehlklassifizierungen kann durch die Veränderung von verschiedenen Variablen (Parameterwahl) beeinflusst werden. Bei der Optimierung des SVM Modells für die Aufgabenstellung verwendet man Trainingsdaten mit bekannter Klassifizierung. Damit entsteht ein Klassifikationsmodell, im Folgenden kurz Modell genannt. Mit Testdaten, die ebenfalls eine bekannte Klassifizierung besitzen, wird ermittelt, wie gut das Modell für die Klassifizierung geeignet ist (siehe Abschnitt2.5. ab Seite 37).

Überanpassungen vermeiden

Liegt eine Überanpassung vor, so ist das gefundene Modell nur scheinbar optimal für die Klassifizierung. Eine Überanpassung an die Testdaten ist nicht wünschenswert, da hierbei nur die Testdaten optimal klassifiziert werden. Daten, die sich von diesen stärker unterscheiden, können jedoch nicht zuverlässig den jeweiligen Klassen zugeordnet werden. Diese Gefahr kann man durch eine Kreuzvalidierung, den Test des Modells mit unterschiedlichen Gruppen von Testdaten, minimieren. Alle Daten mit bekannten Klassifizierungen (siehe *Annotationskorpus* in Abschnitt *2.1.1.*) werden in drei oder mehr Gruppen aufgeteilt. Eine Teilgruppe wird als Testdatenmenge verwendet, alle übrigen werden als eine Trainingsdatenmenge zusammengefasst. Nacheinander werden bei gleicher Parameterwahl alle Teilgruppen einmal als Testdatenmenge verwendet. Diese Methode gibt Aufschluss darüber, wie groß der Grad einer Überanpassung bei dem Modell mit dieser Parameterwahl ist.

2.5. Qualitätsbewertung

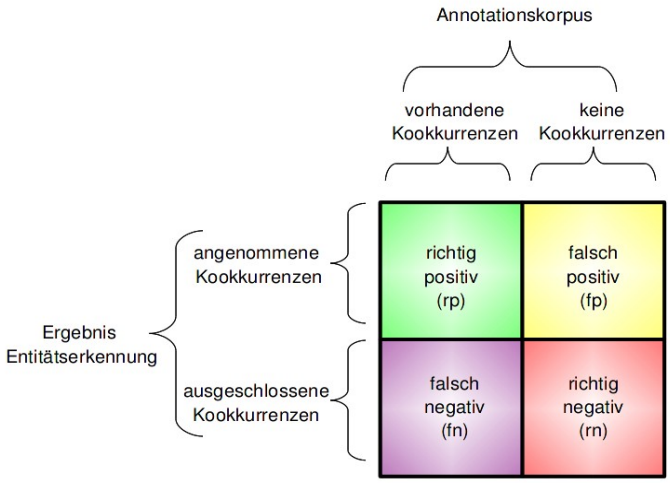
Die Quantität der Ergebnisse einer Identifizierung von Entitäten oder der Klassifizierung ihrer semantischen Beziehungen ist durch Angabe des absoluten Aufkommens leicht messbar, wohingegen eine direkte Bewertung der Qualität der Ergebnisse nicht möglich ist.

Zunächst benötigt man eine Vergleichsmenge, deren Inhalt bekannt und verlässlich ist. Mit einem geeigneten Annotationskorpus kann man verschiedene Kenngrößen für die Qualitätsbewertung errechnen. Sie unterscheiden sich hinsichtlich der Kriterien und ihres Aussagespektrums. Voraussetzung für die hier verwendeten Kenngrößen ist jedoch die vorherige Bestimmung von richtig positiven, falsch positiven, richtig negativen und falsch negativen Ergebnisanteilen. Diese kann man durch den Vergleich der Ergebnisse mit den Einträgen in einem Annotationskorpus bestimmen.

2.5.1. Vergleich mit einem Annotationskorpus

Bei der Bewertung der Ergebnisse der Kookkurrenzsuche werden der Ergebniskorpus (*Kookkurrenzkorpus*) mit dem manuell annotierten Korpus (*Annotationskorpus*) verglichen (Abbildung 2.7).

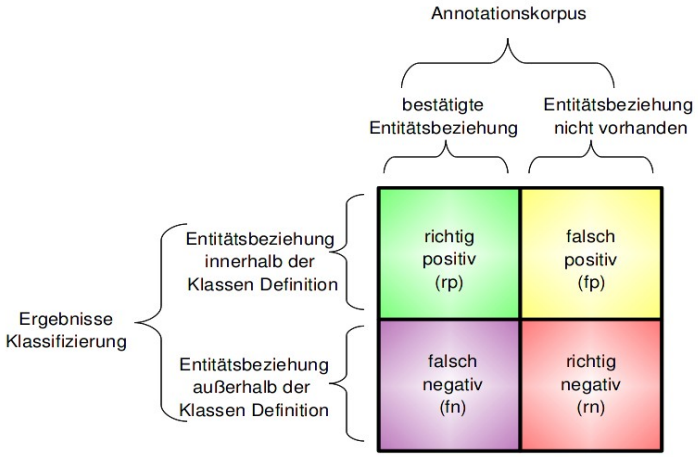
Abbildung 2.7: Eine Wahrheitsmatrix zur Veranschaulichung der Anteile der richtig positiven (rp), richtig negativen (rn), falsch positiven (fp) und falsch negativen (fn) Ergebnisanteile bei der Qualitätsbewertung der Entitätserkennung unter der Bedingung der Kookkurrenz. Die Ergebnisse der kombinierten Entitätserkennung von Enzymen und Krankheiten (angenommene und ausgeschlossene Kookkurrenz) werden mit den manuell annotierten Referenzkorpus (vorhandene und keine Kookkurrenz) verglichen.



Die korrekt erkannte Kookkurrenz einer Krankheits- und Enzymentität wird als richtig positiver (*rp*) Ergebnisanteil, die korrekt erkannte Abwesenheit beider in dem untersuchten Satz, als richtig negativer (*rn*) Ergebnisanteil gewertet. Dem entgegen steht die Anzahl der nicht korrekt vorhergesagten Kookkurrenzen der Entitäten. Ist in der Ergebnismenge eine Kookkurrenz festgehalten, fehlen aber eine oder beide Entitäten in der Annotation des Annotationskorpus, so wird das Ergebnis als falsch positiv (*fp*) bewertet. Für den Fall, dass eine Kookkurrenz der Entitäten im Annotationskorpus vorliegt, jedoch bei der Einordnung des Satzes bei der Entitätserkennung keine Kookkurrenz festgehalten ist, wird das Ergebnis der falsch negativen Menge (*fn*) zugeordnet.

Bei der Bewertung der Leistungsfähigkeit eines Klassifikators werden, ebenso wie bei der Kookkurrenzsuche, die Mengen der richtig positiven (*rp*), richtig negativen (*rn*), falsch positiven (*fp*) und falsch negativen (*fn*) Ergebnisanteile ermittelt (Abbildung 2.8).

Abbildung 2.8: Die Wahrheitsmatrix, die die Grundlage für die Bewertung des Klassifikators bildet. Ebenso wie bei der Kookkurrenzsuche müssen bei der Klassifizierung der semantischen Beziehungen die Mengen der richtig positiven (rp), richtig negativen (rn), falsch positiven (fp) und falsch negativen (fn) Ergebnisanteile durch den Vergleich mit den bestehenden Annotationen in einem Referenzkorpus ermittelt werden.



Die Anteile der richtig positiven (*rp*) und richtig negativen (*rn*) Klassifizierungen der semantischen Beziehung zwischen den beiden Entitäten ergeben sich aus den Fällen, in denen die vorhergesagte Klassifizierung mit der im *Annotationskorpus* festgehaltenen, übereinstimmt. Dem entgegen steht die Anzahl der nicht korrekt vorgenommenen Klassifizierungen der semantischen Beziehungen zwischen Entitäten. Ist die definierte semantische Beziehung zwischen den beiden Entitäten vorhanden, wird jedoch bei der Klassifizierung als 'nicht vorhanden' bewertet, so handelt es sich um ein falsch negatives (*fn),* umgekehrt ein falsch positives (fp) Ergebnis.

2.5.2. Skalare Kenngrößen

Nachdem die Ergebnisse mit dem Annotationskorpus verglichen worden sind, ist es möglich, die Beurteilung der Güte der Ergebnisse unter verschiedenen Gesichtspunkten vorzunehmen.

Die *Präzision* (2.3) ist ein Maß für den Anteil der richtig positiv erkannten Entitäten im Vergleich zu den gesamt als positiv fest gelegten Ergebnissen. *Vollständigkeit* (2.4) gibt Auskunft darüber, wie groß der Anteil der relevanten Ergebnisse ist. Sie sind definiert wie folgt:

$$Präzision = \frac{rp}{rp + fp} \quad (2.3)$$

$$Vollständigkeit = ture\ positive\ rate = \frac{rp}{rp + fn} \quad (2.4)$$

Die *Vollständigkeit* wird auch synonym als *Sensitivität* oder *true positive rate* bezeichnet.

Die Spezifität (2.5) gibt Auskunft über die Rate der richtig negativen Ergebnisanteile gegenüber der Gesamtmenge der negativen Objekte.

$$Spezifität = \frac{rn}{fp + rn} \quad (2.5)$$

Das F_1 *Maß* (2.6) ist der Wert des gewichteten, harmonisierten Mittels zwischen *Präzision* und *Vollständigkeit*.

$$F_1\ Maß = 2 \cdot \frac{Präzision \cdot Vollständigkeit}{Präzision + Vollständigkeit} \quad (2.6)$$

Bei der Optimierung eines Klassifizierers ist es von Vorteil das F_1 *Maß* zu betrachten, da sich bei der Optimierung mit der Erhöhung des Wertes der *Präzision* oft eine Senkung des Wertes der *Vollständigkeit* ergibt. Wünschenswert sind jedoch möglichst hohe Werte beider Kenngrößen. Ein weiterer kombinierter skalarer Wert ist der Korrelationskoeffizient nach Matthews (MCC) [70]:

$$MCC = \frac{rp \times rn - fp \times fn}{\sqrt{(rp + fn)(rp + fp)(rn + fp)(rn + fn)}} \quad (2.7)$$

Der MCC (2.7) variiert zwischen -1 und 1 wobei 0 dem Wert einer Zufallsvorhersage entspricht. Der MCC (2.7) besitzt im Gegensatz zum F_1 Maß (2.6) den Vorteil alle Anteile einer Wahrheitsmatrix zu berücksichtigen, auch den Anteil der richtig negativen Ergebnisanteile und eventuelle unterschiedlich große Mengen der positiven und negativen Ergebnisanteile. Dennoch sollte er mit anderen Kenngrößen kombiniert werden, denn in Fällen, in denen sowohl sehr wenig falsch positive als auch richtig positive Ergebnisanteile erkannt werden, kann er dennoch einen Wert annehmen, der eine gute Vorhersagequalität attestieren würde [71].

Um systematische Fehler zu erfassen, bietet sich die Ermittlung der *Genauigkeit* (2.8, eng. accuracy) an.

$$Genauigkeit = \frac{rp + rn}{rp + rn + fp + fn} \quad (2.8)$$

Die Rate der falsch positiven Einordnungen, *false positive rate* (2.9) wird im Zusammenhang mit den graphischen Darstellungsmöglichkeiten zur Validierung eines Klassifikators wichtig.

$$false\ positive\ rate = \frac{fn}{fp + fn} \quad (2.9)$$

Bei einem *Receiver Operating Characteristics* (ROC) Diagramm (Abbildung 2.9) werden die Vollständigkeit (*true positive rate)* (2.4) und die *false positve rate* (2.9) gegeneinander aufgetragen.

2.5.3. Statistisches Maß der Übereinstimmung

Sofern eine Annotation der gleichen Objekte von mehr als einem Annotator vorgenommen wird, kann man ein statistisches Maß für die Urteilerübereinstimmung bilden. In dieser Arbeit wurde der Koeffizient κ (2.10) nach Cohen [72] verwendet.

$$\kappa = \frac{p_0 - p_z}{1 - p_z} \quad (2.10)$$

Der Anteil p_0 entspricht dem Anteil an gemessener Übereinstimmung und p_z dem Anteil an Übereinstimmung, die zufällig erreicht werden kann. Der Wert für κ kann zwischen 0 und 1 liegen, wobei 0 für eine Übereinstimmung steht und 1 für eine vollständige Übereinstimmung.

2.5.4. Graphische Validierung von Klassifikatoren

Um das Leistungsvermögen eines Klassifikators unter dem Gesichtspunkt verschiedener skalarer Werte zu visualisieren und so leichter vergleichbar zu machen, wurden in dieser Arbeit zwei Arten der graphischen Darstellung gewählt: das Diagramm der *Receiver Operating Characteristics* (ROC) und die Auftragung der Präzision gegen die Vollständigkeit.

Receiver Operating Characteristics

Das *Receiver Operating Characteristics* (ROC) Diagramm wurden in den letzten Jahren vielfach verwendet, um Klassifikatoren, unter anderem basierend auf Methoden des maschinellen Lernens, zu bewerten [73]. Es werden die *true positive rate* (Vollständigkeit) und die *false positive rate* gegeneinander aufgetragen. Die Ausgaben von Klassifikatoren können absolute Werte sein (z.B. -1 = Objekt nicht Teil der Klasse; 1 = Objekt Teil der Klasse) oder aber kontinuierliche Werte im positiven und negativen Zahlenbereich. Die kontinuierlichen Werte können durch einen Schwellenwert gefiltert werden. Ab Erreichen des Schwellenwerts für eine Ausgabe des Klassifikators ist das Objekt Teil einer Klasse, darunter nicht. Variationen des Schwellenwerts führen zu unterschiedlichen Mengenanteilen der klassifizierten Objekte.

Sofern der Klassifikator einem Objekt einen Zahlenwert zuordnet, der mit der Wahrscheinlichkeit korreliert, dass ein Objekt Teil einer definierten Klasse ist, so lässt sich durch eine kontinuierliche Variation des Schwellenwerts ein Graph erzeugen. Aus dem Graphen, der so entsteht, kann man zum einen die generelle Leistungsfähigkeit eines Modells mit anders konfigurierten Klassifikationsmodellen vergleichen (z.B. Vergleich unterschiedlicher Einstellungen von Parametern in einer SVM). Zum anderen kann man für jeden Schwellenwert die Abstimmung von Nutzen (richtig positive Ergebnisanteile) und Kosten (falsch positive Ergebnisanteile) überprüfen [73].

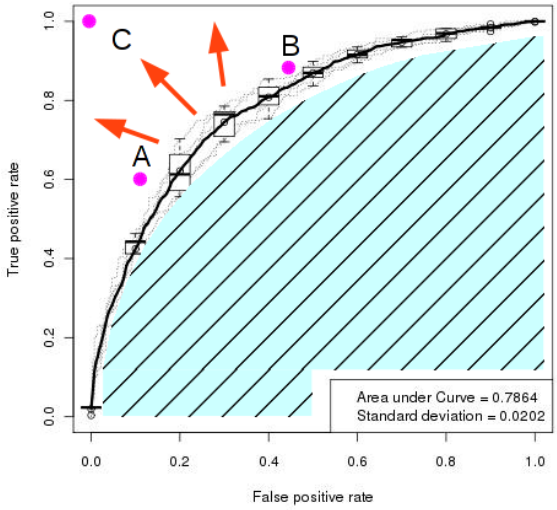


Abbildung 2.9: Exemplarische Abbildung eines Receiver Operator Characteristics (ROC) Graphen. Die schraffierte Fläche skizziert die Fläche deren Größe als Area under Curve (AUC) angegeben wird. Die Punkte A,B und C sind beispielhaft gewählt um die Beschaffenheit bestimmter Klassifikator-einstellungen zu verdeutlichen.

In Abbildung 2.9 ist ein Beispiel für einen ROC Graphen aus der vorgenommenen fünffach Kreuzvalidierung der SVM Klassifizierung gezeigt. Die Fläche unter der Kurve (*AUC*) (Abbildung 2.9, blaue Fläche mit Schraffur) ist als skalarer Wert eine wichtige statistische Größe, die der Wahrscheinlichkeit entspricht, dass ein Klassifikator eine zufällig gewählte positive Instanz höher bewertet als eine zufällig gewählte negative Instanz [73]. Im Allgemeinen wird angestrebt die Fläche unter der Kurve so groß wie möglich werden zu lassen (Abbildung 2.9, Richtung der roten Pfeile). Jedoch werden auch den jeweiligen Positionen bestimmter Punkte auf der Kurve unterschiedliche Gewichtungen zugeordnet. Je näher ein Punkt vom Ursprung der X-Achse entfernt ist und dabei gleichzeitig so weit wie möglich am Ursprung der Y-Koordinate liegt, desto optimaler ist er im Hinblick auf die Kosten-Nutzen Abwägung einer Klassifizierung. In Abbildung 2.9 entspricht das dem *Punkt C*. Eine kostenoptimierte Auswahl der Klassifikatoreinstellungen wäre gegeben, wenn sie eine Ausgabe erzeugen würde, die den Bereich von *Punkt A* Richtung Y-Achse beschreibt. Ein so gewählter Klassifikator nimmt eine positive Klassifizierung nur vor, sofern es eine hoch gewichtete Evidenz dafür gibt. Somit entsteht nur ein kleiner falsch positiver Ergebnisanteil. Ein nutzenoptimiertes Klassifizierungsverhalten wird erreicht, wenn man eine Einstellung wählt, deren Ausgabe im Bereich von *Punkt B* und von dort weiter von der Y-Achse weg liegt. Dort werden so gut wie alle positiven Instanzen richtig erkannt. Das bedingt jedoch einen hohen Anteil an falsch positiven Ergebnissen.

Präzision gegen die Vollständigkeit

Neben der Betrachtung des *ROC* Graphen kann die gleichzeitige Betrachtung einer Auftragung von Präzision und Vollständigkeit sinnvoll sein. So können zwei identische ROC Graphen erhebliche Unterschiede bei der Visualisierung der Präzision und Vollständigkeit zeigen [73]. Deswegen wird begleitend zu den ROC Graphen auch die gleichzeitige graphische Darstellung von Präzision und Vollständigkeit gezeigt.

2.6. Implementation

Den Abschluss in diesem Kapitel bildet eine Erläuterung der technischen Umsetzung der Methoden. Bei der Entwicklungsarbeit der Programmbestandteile wurde grundsätzlich auf eine freie Verfügbarkeit der verwendeten Komponenten für den nichtkommerziellen Gebrauch geachtet. Des Weiteren wurde so weit wie möglich auf die Nutzung fremder Softwarekomponenten verzichtet. Das Ziel war es, den Software-Lebenszyklus der fertigen Anwendungen unter geringem Aufwand zu verlängern, d.h. die Dauer der Verwendbarkeit der Programme zu erhöhen und die Wartbarkeit der Software von Änderungen bei fremden Programmkomponenten weitestgehend zu entkoppeln.

Die Implementierung der Anwendungen, die für die Erstellung der Ergebnisse entwickelt worden sind, liegen in den Programmiersprachen *Python* und *R* sowie in der Skriptsprache *sh-Shell-script* vor. Die Anwendungen wurden in der Betriebssystemumgebung Linux (Debian) ausgeführt. Eine grundsätzliche Möglichkeit der Portierung auf andere Unix Derivate oder auf Microsoft Windows ist nicht ausgeschlossen, wurde aber nicht getestet.

Um die Funktionalität einer SVM zu nutzen, wurde das Programmpaket *SVMlight* [68] eingebunden, welches in der Programmiersprache *C++* erstellt wurde. Es liegt für die Bedienung des Programms eine umfassende Dokumentation vor und das Programmpaket wird nach Abschluss der Entwicklung weiterhin gewartet.

Für die konvertierende Textextraktion in ein Reintextformat wurde das Programm *pdftotext* [74] verwendet.

Die Erstellung der graphischen Darstellung der ROC Kurven sowie der Auftragungen der Präzision gegen die Vollständigkeit (siehe Graphische Validierung von Klassifikatoren auf Seite 42) erfolgte mit Hilfe des *R* Pakets ROCR [75].

Die Zwischenergebnisse und Endresultate der Berechnungen wurden in Datenbanken eines MySQL Servers abgelegt. Dadurch war es möglich während der Verarbeitung effizient auf benötigte Daten zugreifen zu können und die Ergebnisse abschließend zu archivieren. Eine Übersicht über die gesamte Programmausrüstung findet sich im Anhang am Ende dieser Arbeit.

3. Ergebnisse und Diskussion

3.1. Wissensquelle Text

Das zu Grunde liegende Ausgangsmaterial für alle Ergebnisse dieser Arbeit sind publizierte Artikel aus dem Bereich der Lebenswissenschaften sowie deren Titel und Kurzzusammenfassungen. Eine der umfassendsten Sammlungen dieser Art Quellmaterials ist die PubMed. Im folgenden Abschnitt werden die PubMed Titel und Kurzzusammenfassungen im Hinblick auf ihre Wortzusammensetzung betrachtet sowie der Inhalt der für die Entitätserkennung und die Klassifizierung verwendeten Wörterbücher beschrieben und diskutiert.

3.1.1. Wörter in PubMed Titeln und Kurzzusammenfassungen

Es wurde ein Wörterbuch aller Wörter, der in der PubMed Datenbank vorkommenden Titel und Kurzzusammenfassungen erstellt. Es entsprach dem Stand der PubMed Datenbank vom Dezember 2009. Eine Analyse der Titel und Kurzzusammenfassungen ergab, dass in den 19.254.163 betrachteten PubMed Einträgen insgesamt über 2 Milliarden Einzelwörter enthalten sind (Tabelle 3.1). Die Gesamtmenge der Einzelwörter beträgt 13.6 Millionen (Tabelle 3.1) und lässt darauf schließen, dass die unterschiedlichen Einzelwörter überwiegend mehrfach vorkommen. Etwa 11 Mio. Einträge in PubMed enthalten eine Kurzzusammenfassung und somit enthält eine Kurzzusammenfassung einer PubMed Referenz durchschnittlich 168 Wörter.

Tabelle 3.1 : Wörter in Titeln und Kurzzusammenfassungen der PubMed Datenbank

	Anzahl
Referenzen	19.254.163
Anzahl aller Wörter in Titeln	213.815.565
Unterschiedliche Wörter in Titeln	2.797.741
Anzahl aller Wörter in Kurzzusammenfassungen	1.857.691.989
Unterschiedliche Wörter in Kurzzusammenfassungen	12.511.460
Gesamtanzahl aller Wörter	2.071.507.554
Gesamtanzahl der unterschiedlichen Wörter	13.611.924

Das erstellte Wörterbuch enthält alle Worte als Reintext und als Hashwertrepräsentation (*2.2.1. Hashwerte statt natürliche Sprache* auf Seite 22). Zum einen diente dieses Wörterbuch der Abschätzung der Anzahl der Dimensionen, der von der SVM zu

verarbeitenden Vektoren. Die Anzahl der Dimensionen wird bestimmt durch die Anzahl der unterschiedlichen Wörter im *Klassifikationskorpus* (*2.2.3. Aufbereitung und Termgewichtung* auf Seite 24). Der *Klassifikationskorpus* ist eine Teilmenge der PubMed Titel und Kurzzusammenfassungen und kann maximal so viele unterschiedliche Wörter beinhalten, wie die PubMed Titel und Kurzzusammenfassungen. Zum anderen ermöglichte das Wörterbuch die Überführung des *Annotationskorpus* in eine Form, die für eine manuelle Annotation der darin enthaltenen Titel und Sätze notwendig war. Der *Annotationskorpus* besteht ebenfalls aus einer Teilmenge der PubMed Titel und Kurzzusammenfassungen. Durch das Wörterbuch konnte jedes Wort des *Annotationskorpus*, das als Hashwert vorlag, in seine Reintextform für die manuelle Annotation überführt werden.

Stoppwörter

In Tabelle 3.2 sind die 25 am häufigsten auftretenden Wörter in PubMed Titeln und Kurzzusammenfassungen aufgeführt. Bis auf das Wort „patients“ handelt es sich dabei um sogenannte Stoppwörter, also Wörter mit vermeintlich niedrigem Informationsgehalt (*2.1.2 Stoppwortlisten* auf Seite 21). Die Stoppwortlisten wurden im Zusammenhang der Dimensionsreduktion für die Klassifizierungseingaben der SVM getestet. Eine Aufstellung aller als Stoppwort betrachteten Wörter mit der Information über die Häufigkeit ihres Vorkommens findet sich in *Anhang B. Liste der verwendeten statischen Stoppwörter* auf Seite 95.

Tabelle 3.2: Übersicht der 25 häufigsten Wörter in PubMed Titeln und Kurzzusammenfassungen. Neben dem Wort ist die Häufigkeit des Vorkommens aufgeführt.

Wort	***Anzahl ihres Mehrfachvorkommens***
the	115.223.746
of	102.361.674
and	69.880.428
in	62.403.144
to	36.675.053
a	34.463.386
with	24.890.653
was	18.612.934
for	18.565.417
were	16.179.616
by	14.878.072
is	13.924.580
that	13.872.233
on	10.969.312
from	9.544.213
patients	9.395.163
as	9.242.925
this	8.114.998
or	7.953.376
at	7.597.546

Die ursprüngliche Stoppwortliste[5], wurde nach der Analyse der Wortfrequenz in den PubMed Kurzzusammenfassungen angeglichen. Des Weiteren wurde eine zweite Stoppwortliste aus einer Teilmenge der ersten Stoppwortliste gebildet. In der zweiten Liste wurden die Wörter als Stoppwort ausgeschlossen, die für die semantische Bedeutung eines Satzes wichtig erschienen, wie z.b. „not“ mit einer Häufigkeit von 5,5 Millionen oder „causes“ mit einer Häufigkeit von mehr als 0,3 Millionen in den PubMed Titeln und Kurzzusammenfassungen.

Kurzzusammenfassung versus Volltext

Die biomedizinische Publikation ist immer noch *das* Medium, in dem neue Erkenntnisse einer interessierten Fachöffentlichkeit vorgestellt werden. Dies zeigt sich auch durch den Umstand, dass biomedizinische Publikationen einen Zuwachs im doppelt exponentiellen Bereich verzeichnen [76]. Die Kollektion der Referenzen in der PubMed

[5] ftp://ftp.cs.cornell.edu/pub/smart/english.stop

umfasst mittlerweile mehr als 20 Millionen Einträge (Stand Juni 2011) und allein die darin enthaltenen Referenzen, die im Jahr 2010 veröffentlicht und hinzugefügt wurden, betragen über 918.000. Mehr als 11 Millionen PubMed Einträge enthalten eine Kurzzusammenfassung. Autoren von Artikeln sind meist darum bemüht die zentralen Inhalte und Erkenntnisse in einer Kurzzusammenfassung zum Ausdruck zu bringen. Trotzdem zeigen aktuelle Untersuchungen, dass die Volltextsuche die Wahrscheinlichkeit erhöht einen relevanten Artikel zu finden [77]. Mit den heutigen Mitteln des Text und Data Minings sind Kurzzusammenfassungen noch eine adäquatere Wahl als Quelle für die Verarbeitung. Dies gilt im Hinblick sowohl auf die Präzision als auch auf die Verarbeitungszeit, sofern es sich um semantische Fragestellungen handelt, da sie aufgrund ihrer simpleren Struktur im Vergleich zum Volltext leichter zu erfassen sind [78]. Dennoch zeigt sich, dass die Suche nach bestimmten Entitäten im Volltext von bezüglich der Anzahl mehr Ergebnisse erbringt verglichen mit der Suche in Kurzzusammenfassungen [78].

3.1.2. Wörterbücher der Krankheiten und Enzyme

Krankheitswörterbuch

Für die Suche nach Krankheitsentitäten wurde ein Wörterbuch aus den Begriffen für Krankheiten und pathologische Zustände des MeSH Thesaurus gebildet. Das Krankheitswörterbuch wurde für zwei Aufgaben benötigt: Die Identifizierung von Krankheitsentitäten im Kookkurrenzkorpus (*3.3. Gemeinsam auftretende Krankheiten und Enzyme* auf Seite 56) und zur Vorverarbeitung des *Klassifikationskorpus* (*2.2.3. Aufbereitung und Termgewichtung* auf Seite 24). Das für die Suche nach Krankheitsentitäten generierte Wörterbuch umfasst 22.380 Begriffe zu 4.054 Krankheitsentitäten. Es enthält damit durchschnittlich zwischen fünf und sechs Synonyme für eine Krankheitsentität. Die 15 Krankheiten mit der größten Anzahl an Synonymen sind in Tabelle 3.3 aufgeführt.

Tabelle 3.3: Eine Übersicht über die 15 Krankheiten und pathologischen Zustände mit den meisten Synonymen im Krankheitswörterbuch. Die Namen der Krankheiten und pathologischen Zustände sind jeweils in deutscher und englischer Sprache angegeben.

Deutsch	**Englisch**	**Anzahl Synonyme**
Meningeom	Meningioma	54
Non-Hodgkin Lymphom	Non-Hodgkin Lymphoma	48
Tremor	Tremor	48
Aphasie	Aphasia	46
Mukolipidose	Mucolipidoses	45
Erkrankungen der Pupille	Pupil Disorders	41
Agnosie	Agnosia	41
Halluzinationen	Hallucinations	39
Primäre und sekundäre Neoplasien	Primary and secondary Neoplasms	37
Verkrampfungen/Krampfanfälle	Seizures	37
Dystonien	Dystonic Disorders	36
Neoplasien der Gehirnregion	Brain Neoplasms	36
Epilepsie	Epilepsy	35
Neoplasien des Hirnstamms	Brain Stem Neoplasms	35
Kopfschmerz	Headache	34

Der MeSH Thesaurus ist hierarchisch aufgebaut und so können gezielt Begriffe extrahiert werden, die Krankheiten umfassen und bestimmte unerwünschte Begriffe (Krankheiten aus der Veterinärmedizin) ausgeschlossen werden (zur MeSH Begriffsauswahl siehe auch Tabelle 2.1 auf Seite 19). Darüber hinaus wird die MeSH Sammlung ständig manuell von Experten gepflegt und aktualisiert [79]. MeSH ist mit 26.000 [6] Begriffen nicht so umfangreich, wie die UMLS Ontologie (*1.3 Biologische und medizinische Wissenssammlungen* auf Seite 7) mit etwa 730,000 Konzepten [28] aus dem biomedizinischen Bereich. Allerdings ist der Aufwand für die Erstellung und die Pflege eines Wörterbuchs, das der präzisen Suche nach Krankheitsentitäten dienen soll, mit der Quellsammlung UMLS wesentlich aufwändiger. Die UMLS Ontologie ist eine Sammlung aus unterschiedlichen Kollektionen von Fachvokabular, und die einzelnen Elemente unterliegen zum Teil unterschiedlichen Lizenzierungen. Die Überprüfung der Qualität der einzelnen Fachvokabulare ist nicht ohne erheblichen Mehraufwand möglich, denn die Kontrolle auf falsche Synonyme und nicht differenzierbare Homonyme, müsste bei jeder Aktualisierung durchgeführt werden.

[6] Fact Sheet Medical Subject Headings 2011

Enzymwörterbuch

Das Enzymwörterbuch wurde aus der BRENDA Datenbank Aktualisierungsroutine [46] übernommen. Dieses wird dort bei jeder halbjährlichen Aktualisierung der Datenbank neu gebildet. Das Enzymwörterbuch, das bei der Vorverarbeitung des *Klassifikationskorpus* (*2.2.3. Aufbereitung und Termgewichtung* auf Seite 24) für die in dieser Arbeit vorgestellten Ergebnisse der Klassifizierung (*3.4. Die Klassifizierten von Entitätsbeziehungen* auf Seite 64) verwendet wurde, enthielt 58.646 Begriffe zu 4.618 Enzymentitäten. Alle Enzymnamen, die in das Wörterbuch einfließen, entstammen der BRENDA und wurden manuell aus wissenschaftlichen Publikationen extrahiert oder entspringen der IUBMB empfohlenen und systematischen Benennung [44]. In BRENDA sind für ein Enzym durchschnittlich 15 Synonyme bekannt [43]. Bei der Bildung des Enzymwörterbuchs für die Verwendung bei der Enzymentitätensuche im Rahmen der BRENDA Datenbank Aktualisierungsroutine werden nicht differenzierbare Synonyme, die auch Homonyme für andere Entitäten sein könnten, entfernt. Damit ist es eine qualitativ hochwertige und umfassende Ressource zur Verwendung bei Verarbeitungsroutinen, die die Kenntnis von Enzymnamen bedingen.

3.2. Betrachtung des Annotationskorpus

Der *Annotationskorpus* enthält 5031 Sätze. Er wurde zur Validierung der Ergebnisse und für das Trainieren und Testen der SVM Klassifizierung zusammengestellt. Die Filterung zur Auswahl der darin enthaltenen Titel und Sätze erfolgte unter der Annahme des darin gehäuften Vorkommens von Enzymen und Krankheiten (*2.1.1 Annotationskorpus* auf Seite 15). Die manuelle Annotation ergab, dass in 4.543 Sätzen mindestens ein Enzym und in 2.441 Sätzen mindestens eine Krankheit präsent war. In 2.176 Sätzen wurde eine Kookkurrenz von Enzym und Krankheit bestätigt. In Tabelle 3.4 ist die bei der Zusammenstellung angenommene und die durch die Annotation bestätigte Anzahl der vorkommenden Enzyme und Krankheiten aufgelistet.

Tabelle 3.4: Vergleich der Anzahl des Auftretens von Enzymen und Krankheiten in den Sätzen und Titeln des Annotationskorpus. Die Anzahl für das angenommene Auftreten ergibt sich aus den Vorbedingungen für die Zusammenstellung (2.1.1 Annotationskorpus auf Seite 15). Die Anzahl für das bestätigte Auftreten ergab sich nach der durchgeführten manuellen Annotation.

	Anzahl
Angenommenes Auftreten von Enzymen	4.500
Bestätigtes Auftreten von Enzymen	4.543
Angenommenes Auftreten von Krankheiten	2.500
Bestätigtes Auftreten von Krankheiten	2.441
Sätze mit Kookkurrenzen	2.176

Anteile der semantischen Relationen

Alle 2.176 Sätze und Titel, in denen mindestens eine Kookkurrenz eines Enzyms und einer Krankheit festgehalten wurde, sind auch auf die semantische Relation untereinander untersucht worden. Die semantischen Relationen wurden festgehalten, sofern sie einer oder mehreren der zuvor definierten Kategorien entsprach (*2.4.1. Definitionen der Entitätsrelationen* auf Seite 33). Die Anzahl der gefunden Sätze und Titel, in denen die festgestellte Relation den definierten Kategorien entsprach, ist in Tabelle 3.5 aufgeführt.

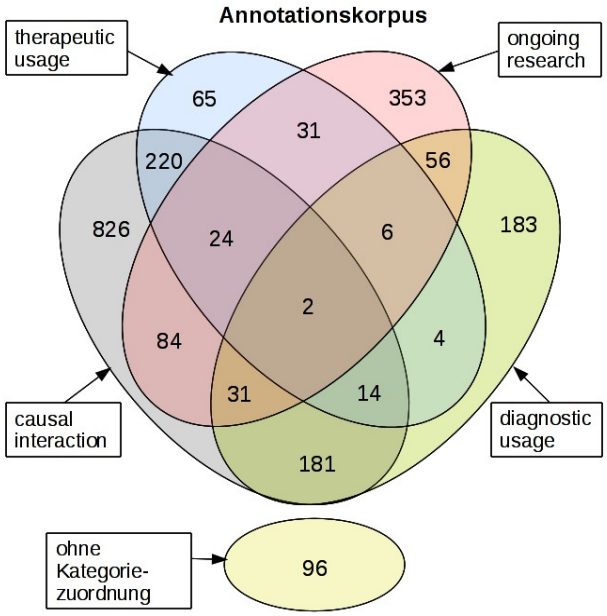
Tabelle 3.5: Eine Übersicht der Anzahl der auftretenden Beziehungen von Enzymen und Krankheiten in den Sätzen und Titeln des Annotationskorpus, die den definierten Kategorien für die Entitätsrelationen entsprechen.

Kategorie	**Anzahl**
Causal interaction	1.382
Ongoing research	587
Diagnostic usage	477
Therapeutic application	366
Gesamtmenge	2.812

Ein Überblick über die Größe der Schnitt- und Differenzmengen der jeweiligen Kategorien ist in Abbildung 3.1 graphisch dargestellt. Die Kategorie *diagnostic usage* (Abbildung 3.1, rosa) hat, verglichen zur Gesamtmenge von 477 festgestellten

Beziehungen, die größte sich nicht überschneidende (unabhängige) Menge. Die Kategorie *therapeutic application* (Abbildung 3.1, blau) hat, verglichen zur Gesamtmenge von 366 festgestellten Beziehungen, die anteilig größte Schnittmenge mit einer anderen Kategorie, nämlich *causal interaction*.

Abbildung 3.1: Die Verteilung der Zuordnung zu semantischen Relationen im Annotationskorpus.



Für diese Betrachtung wurden alle 2.176 Sätze und Titel im Annotationskorpus einbezogen, in denen eine Kookkurrenz einer Enzym- und Krankheitsentität festgestellt wurde. Angegeben ist die Größe der Schnitt- und Differenzmengen der Sätze und Titel im Annotationskorpus, deren semantische Relation von Enzym- und Krankheitsentität jeweils einer oder mehreren Kategorien causal interaction (grau), therapeutic application (blau), ongoing research (rosa) und diagnostic usage (grün) zugeordnet wurden. Die Anzahl der Kookkurrenzen, die keiner Kategorie zugeordnet wurden, ist unterhalb in dem gelben Oval angegeben.

Diversität der Interpretation einer Relation

Um den Grad an Übereinstimmung bei einer kategorisierenden Zuordnung zu bestimmen und den Anteil des Einflusses der individuellen Interpretation auf die inhaltliche Analyse der Texte des *Annotationskorpus* festzustellen, wurde die Zuordnung durch eine andere Person wiederholt. Es wurden 2139 Sätze und Titel des *Annotationskorpus* nochmals annotiert [80]. In diesen Sätzen und Titeln lag mindestens

eine Kookkurrenz eines Enzyms und einer Krankheit vor und sie wurden ohne Kenntnis der Erstannotation, den Kategorien *causal interaction*, *therapeutic application*, *ongoing research* und *diagnostic usage* zugeordnet. Der Anteil der Übereinstimmungen für alle überprüften Sätze und Titel variierte zwischen 89.6% für die Kategorie *therapeutic application* und 66,0% bei der Kategorie *ongoing research*.

Eine Auswertung anhand des statistischen Werts des Koeffizienten κ für die Urteilerübereinstimmung nach Cohen [72] ist in Tabelle 3.6 aufgeführt (*2.5.3. Statistisches Maß der Übereinstimmung* auf Seite 41).

Tabelle 3.6: Errechnete Urteilerübereinstimmung für die Zuordnung von Kookkurrenzen zu den Klassifizierungskategorien.

Kategorie	**κ**
therapeutic application	0,586
causal interaction	0,302
diagnostic usage	0,248
ongoing research	0,243

Die Übereinstimmung ist bei der Kategorie *therapeutic application* am größten. Für alle anderen Kategorien ist das Maß der Übereinstimmung geringer. Bei nachträglicher Betrachtung einiger Sätze, bei denen die Annotatoren nicht übereinstimmten, konnte festgestellt werden, dass hier oft besonders lange, vage oder verklausulierte Formulierungen vorlagen oder es sich um Mehrfachkookkurrenzen unterschiedlicher Entitäten handelte. Solche Formulierungen lassen einen Spielraum bei der Zuordnung der Kookkurrenzen zu, der dann durch die individuelle Interpretation der Annotatoren zu einem unterschiedlichen Maß der Übereinstimmung führt. Die erreichten κ-Werte sind zwar erniedrigt (Tabelle 3.6), jedoch wird in der Linguistik die Berechnung der Urteilerübereinstimmung nach Cohen für die inhaltliche Analyse von Texten kritisch gesehen [81] und es wird empfohlen die Beurteilungsschwelle der κ-Werte entsprechend zu relativieren [82].

Öffentliche Referenztextkörper

Mittlerweile stehen einige biomedizinische Textkörper zum Test von Text Mining Anwendungen zur Verfügung, z.b. der GENIA [83] Korpus. Jedoch konnte keiner der für diese Arbeit in Betracht gezogenen Referenztextkörper Verwendung finden, weil er entweder Enzym- und Krankheitsentitäten nicht in Kombination in einem Korpus annotiert enthielt [84], oder Enzym- und Krankheitsentitäten nicht in ausreichender

Menge aufwies [85], um statistisch aussagekräftige Vergleiche machen zu können. In allen diesen Textkörpern fehlte eine Einteilung entsprechend der in dieser Arbeit definierten Kategorien, die auf jeden Fall selbst vorgenommen werden musste.

3.3. Gemeinsam auftretende Krankheiten und Enzyme

Im Rahmen der halbjährlichen Aktualisierung der BRENDA Datenbank wird durch eine automatische Auswertung der PubMed Titel und Kurzzusammenfassungen unter anderem nach dem Auftreten von Enzymentitäten gesucht. Die Ergebnisse dieser Suche wurden für die Identifizierung von Kookkurrenzen mit Krankheitsentitäten verwendet (siehe *genutzten Ressourcen des BRENDA Informationssystems* in Abschnitt 2.3.2 auf 30). Um das gemeinsame Auftreten mit Krankheiten festzustellen, wurden die Krankheitsentitäten unter Verwendung des Krankheitswörterbuchs in den PubMed Titeln und Kurzzusammenfassungen gesucht. Das Verfahren der Kookkurrenzsuche von Enzym- und Krankheitsentitäten ist seit der Aktualisierung 2009/2 Bestandteil der BRENDA Aktualisierungsroutine [43]. Die folgenden Angaben beziehen sich, sofern nicht anders angegeben, auf die Ergebnisse, die der Aktualisierung 2010/2 entnommen wurden.

Enzyme und Krankheiten in Kookkurrenz

Es konnten 4.734 unterschiedliche Krankheiten gefunden werden. Davon traten 4.061 Krankheiten gemeinsam mit einem Enzym in einem Satz oder Titel auf. In den Ergebnissen der Kookkurrenzsuche wurden ein gemeinsames Auftreten mit 1.979 unterschiedlichen EC Nummern in 522.720 Referenzen festgehalten (Tabelle 3.7). Es wurden Kookkurrenzen von Enzym- und Krankheitsentitäten in 624.836 Sätzen von Kurzzusammenfassungen und in 187.601 Titeln gefunden.

Tabelle 3.7: Die Ergebnisse der Kookkurrenzsuche (Juli 2010) [43]. Es ist die Anzahl der unterschiedlichen Referenzen, Krankheiten und EC Nummern sowie die Anzahl der unterschiedlichen Kombinationen, die dafür auftraten, angegeben. Begleitend dazu sind die ermittelten Werte für Präzision, Vollständigkeit und das F_1 Maß aufgeführt.

Aktualisierungen	**Juli 2010**
Kombination	910.897
Referenzen	522.720
Krankheiten	4.061
EC Nummern	1.979
Präzision	0,835
Vollständigkeit	0,948
F_1 Maß	0,888

Die häufigsten Krankheiten

In der Tabelle 3.8 sind die zehn Krankheiten aufgeführt, die am häufigsten in den Ergebnissen der Kookkurrenzsuche auftraten. Die mit Abstand am häufigsten gefundenen Krankheiten sind Erkrankungen mit Tumorbildung (Neoplasien). Ebenso ist die Anzahl der assoziierten EC Nummern mit Abstand die höchste (Tabelle 3.8). Neoplasien treten in Referenzen doppelt so häufig auf, wie die am zweithäufigsten gefundenen Krankheiten, den Erkrankungen der Herzkranzgefäße. Die Erkrankungen der Herzkranzgefäße sind wiederum auch doppelt so häufig vertreten, wie die nachfolgende Gruppe der Infektionen. Wohingegen die Anzahl der, mit Infektionen assoziierten EC Nummern weitaus höher ist, als bei den Erkrankungen der Herzkranzgefäße.

Tabelle 3.8: Eine Übersicht über die zehn Krankheiten und pathologischen Zustände, die am häufigsten zusammen mit Enzymen gefunden wurden. Die Namen der Krankheiten und pathologischen Zustände sind jeweils in Deutsch und Englisch angegeben. Begleitend ist die Anzahl der Referenzen, in denen sie gefunden wurden und die Anzahl der unterschiedlichen EC Nummern aufgeführt, die in den Ergebnissen der Kookkurrenzsuche mit diesen Krankheiten assoziiert auftraten.

Deutsch	Englisch	Referenzen	EC Nummern
Neoplasien	Neoplasms	83.527	1.175
Erkrankungen der Herzkranzgefäße	Coronary Artery Disease	44.939	185
Infektionen	Infection	21.678	955
Wunden und Verletzungen	Wounds and Injuries	21.656	743
Karzinome	Carcinoma	17.948	764
Bluthochdruck	Hypertension	14.062	387
Myokardiale Infarkte	Myocardial Infarction	13.334	320
Schmerzen	Pain	12.144	374
Neoplasien der Brust	Breast Neoplasms	11.253	584
Cholera	Cholera	8.932	229

Die häufigsten Enzyme

In Tabelle 3.9 sind die zehn EC Nummern aufgeführt, die am häufigsten in den Ergebnissen der Kookkurrenzsuche auftraten. Die *cGMP-abhängige Proteinkinase* (EC 2.7.11.12) ist das Enzym, das am häufigsten in den Ergebnissen zu finden ist. Es tritt mehr als dreimal so häufig auf, als das zweithäufigste Enzym, die *Rezeptor-Tyrosinkinase* (EC 2.7.10.1), welche die Phosphorylierung von Tyrosin-Resten katalysiert. Das Angiotensin-konvertierende Enzym (EC 3.4.15.1, empfohlener Name peptidyl-dipeptidase A), das an dritter Stelle in Tabelle 3.9 steht, ist an der Regulation des Blutdruckes und des Elektrolyt-Wasserhaushalts beteiligt und ist das Wirkziel der Medikamentengruppe der ACE (Angiotensin Converting Enzyme)-Hemmer. ACE-Hemmer werden häufig bei therapeutischen Interventionen gegen Bluthochdruck, Herzinsuffizienz, Herzinfarkt, Herzmuskelentzündung und diabetischer Nephropathie eingesetzt.

Tabelle 3.9: Eine Übersicht über die zehn EC Nummern, die am häufigsten zusammen mit Krankheiten gefunden wurden. Die EC Nummern sind jeweils begleitend mit ihren durch die IUBMB empfohlenen Namen angegeben sowie die Anzahl der Referenzen, in denen sie gefunden wurden und die Anzahl der Krankheiten, die in den Ergebnissen der Kookkurrenzsuche mit diesen EC Nummern assoziiert auftraten.

EC Nummer	Empfohlener Name	Referenzen	Krankheiten
2.7.11.12	cGMP-dependent protein kinase	66.226	1.049
2.7.10.1	Receptor protein-tyrosine kinase	20.053	997
3.4.15.1	peptidyl-dipeptidase A	10.574	661
3.4.23.49	omptin	10.490	980
3.2.2.22	rRNA N-glycosylase	10.220	349
3.4.21.53	Endopeptidase La	9.488	951
1.14.99.1	Prostaglandin-endoperoxide synthase	9.377	832
1.9.3.1	cytochrome-c oxidase	9.109	909
2.7.7.49	RNA-directed DNA polymerase	8.240	794
3.1.1.7	Acetylcholinesterase	8.237	549

Die häufigsten Kombinationen von Krankheiten und Enzymen

Es konnte ermittelt werden, dass bestimmte Erkrankungen häufiger zusammen mit bestimmten EC Nummern genannt werden. In Tabelle 3.10 sind die fünfzehn Krankheiten aufgeführt, die am häufigsten mit einem bestimmten Enzym in den Ergebnissen verknüpft waren. Es könnte darauf hindeuten, dass diese Krankheiten in besonderer Verbindung zu den Enzymen stehen. Die allgemeinen Erkrankungen der Herzkranzgefäße treten in Kombination mit der *cGMP-abhängigen Proteinkinase* am häufigsten in den Ergebnissen auf (Tabelle 3.10). Ebenso wie zum kardialen Erkrankungsspektrum gehörenden myokardiale Infarkte, myokardiale Ischämien und koronare Stenosen die auch unter den zehn häufigsten Kombinationen zu finden sind. Die *cGMP-abhängige Proteinkinase* ist ebenso in der Rangfolge der am häufigsten gefundenen Enzyme in Tabelle 3.9 das mit Abstand am häufigsten vertretene.

Tabelle 3.10: Eine Übersicht über die zehn Krankheiten, die am häufigsten mit einer bestimmten EC Nummer gefunden wurden. Die EC Nummern sind jeweils begleitend mit ihren durch die IUBMB empfohlenen Namen und der gemeinsam gefundenen Krankheit angegeben sowie zusammen mit der Häufigkeit der Nennung der Krankheit mit diesem Enzym in den Ergebnissen der Kookkurrenzsuche.

EC Nummer	empfohlener Name	Krankheit	Referenzen
2.7.11.12	cGMP-dependent protein kinase	Erkrankungen der Herzkranzgefäße	44.787
2.7.10.1	Receptor protein-tyrosine kinase	Neoplasien	8.647
3.2.2.22	rRNA N-glycosylase	Cholera	8.361
2.7.11.12	cGMP-dependent protein kinase	Myokardiale Infarkte	7.281
3.1.1.7	Acetylcholinesterase	Schmerzen	6.773
3.4.21.77	Semenogelase	Neoplasien der Prostata	4.352
3.4.23.15	Renin	Bluthochdruck	3.628
2.7.11.12	cGMP-dependent protein kinase	Koronare Stenosen	3.404
2.7.7.49	RNA-directed DNA polymerase	Neoplasien	3.293
3.6.3.49	channel-conductance-controlling ATPase	Zystische Fibrose	3.222
1.14.99.1	Prostaglandin-endoperoxide synthase	Neoplasien	3.193
2.7.11.12	cGMP-dependent protein kinase	Bluthochdruck	2.855
3.2.1.49	alpha-N-acetylgalactosaminidase	Humane Influenza	2.815
3.4.21.97	Assemblin	Infektionen	2.739
2.7.11.12	cGMP-dependent protein kinase	Myokardiale Ischämie	2.510

Die cGMP-abhängige Proteinkinase in der Literatur

Es zeigte sich eine auffällige Häufung bei den Ergebnissen der Kookkurrenzsuche, die eine Verbindung zwischen dem kardiovaskulären System und der *cGMP-abhängigen Proteinkinase* vermuten lässt. Diese Verbindung wird in medizinischen Publikationen zur *cGMP-abhängigen Proteinkinase* bestätigt. Die *cGMP-abhängige Proteinkinase* ist in Signalwege eingebunden, die sich umfassend auf das kardiovaskuläre System auswirken [86]. Es handelt sich dabei um Signalwege, die der arteriellen Hypertonie, der Volumenüberlastung sowie der kardialen Hypertrophie und Insuffizienz entgegenwirken [86]. Darüber hinaus wird die *cGMP-abhängige Proteinkinase* als viel versprechendes Wirkziel für neue therapeutische Ansätze in diesem Zusammenhang betrachtet [87].

Anteile der unterschiedlichen Enzymklassen

Die Verteilung der Ergebnisse nach Enzymklassen ist in Tabelle 3.11 aufgeschlüsselt. Die Enzymklasse der Hydrolasen ist im Vergleich zu den anderen Enzymklassen überrepräsentiert, die Klasse der Ligasen wiederum unterrepräsentiert. Obwohl die

Isomerasen nur 186 unterschiedliche EC Nummern in BRENDA umfassen sind sie doppelt so häufig repräsentiert (0,86) wie die Ligasen (0,45), die 161 EC Nummern umfassen.

Tabelle 3.11: Ein Überblick über die Verteilung der Ergebnisse der Kookkurrenzsuche (2010/2) von Enzymen und Krankheiten.

Enzymklassen Name	EC	Krankheiten* mit einer Enzym-kookkurrenz	EC Nummern* mit einer Krankheits-kookkurrenz	Kookkurrenzen* von EC Nummer und Krankheit (a)	Kombinationen* von EC Nummer, Krankheit und Referenz	EC Nummern* in BRENDA (b)	Repräsentation a/b
Oxidoreduktasen	1	2.646	429	23.214	159.215	1.393	0,76
Transferasen	2	2.862	506	23.939	283.794	1.369	0,78
Hydrolasen	3	3.439	737	55.036	406.831	1.523	1,64
Lyasen	4	1.721	169	6.053	31.487	494	0,56
Isomerasen	5	1.350	64	3.531	22.230	186	0,86
Ligasen	6	712	74	1.578	7.340	161	0,45
	1-6	4.061	1.979	112.805	910.897	5.126	

**) Angegeben ist die Anzahl jeweils unterschiedlicher Elemente.*

Wie in Tabelle 3.11 gezeigt, sind die Anteile von Enzymen und Krankheiten nicht gleichmäßig über alle Enzymklassen verteilt. Die Klasse der Hydrolasen (EC 3) wird häufiger in Verbindungen mit pathologischen Vorgängen gefunden und stellt einen Anteil von fast 50% der Kombinationen aus EC Nummer/Krankheit/Referenz. Hervorzuheben ist die Sub-Klasse der Peptidasen (EC Sub-Klasse 3.4.), die die Spaltung von Peptiden und Proteinen katalysieren und 53% der Kombinationen aus EC Nummer/Krankheit/Referenz der Hydrolasen ausmachen.

Die Dominanz der Peptidasen

Die Peptidasen stellen die Sub-Klasse mit den meisten offiziellen EC Nummern und repräsentieren mit 41% einen beachtlichen Anteil der bislang in der EC-Klassifikation aufgenommenen Hydrolasen. Die Dominanz der Peptidasen unter den Hydrolasen erklärt sich aber nicht nur durch ihre schiere Vielzahl unterschiedlicher Enzyme. Peptidasen sind katalytisch an vielen essentiellen Schlüsselstellen im eukaryotischen Leben aktiv, wie z.B. bei der Gerinnungskaskade (Thrombin, Plasmin), dem immunologischen Komplementsystem (Komplement Protease C1r, Komplement Faktor D) und der Verdauung des Proteinanteils der Nahrung (Pepsin, Trypsin). Infolge dessen

kann der funktionelle Ausfall einer Peptidase eine Reihe pathologischer Prozesse in Gang setzen. Als bakterielle Virulenzfaktoren sind voll funktionstüchtige Peptidasen ebenfalls eine Gefahr für den Organismus, so z.B. die von *Bacillus anthracis* sekretierte Endopetidase des Anthrax Toxins. In BRENDA stehen über 20% der Einträge in Verbindung mit Inhibitoren im Zusammenhang mit Peptidasen. Dies weist auf ein hohes Forschungsinteresse hin, besonders im medizinischen Bereich. Denn die Inhibition eines proteolytischen Enzyms wird oft als potentieller Teil einer therapeutischen Strategie in Betracht gezogen. Im Vergleich haben andere Enzymklassen nicht so dominant hohe Einträge in Verbindung mit Inhibitoren. Die Ligasen (EC 6) sind eher weniger mit Krankheiten verbunden (Tabelle 3.11) und stellen auch nur einen Anteil von 3% bei den Einträgen zu Inhibitoren.

Fortschritte in der Kookkurrenzsuche

Die im Rahmen dieser Arbeit implementierte Kookkurrenzsuche ist erstmalig bei der BRENDA Aktualisierung 2009/2 einbezogen worden. Seitdem sind fünf BRENDA Aktualisierungen durchlaufen worden. In Tabelle 3.12 sind die Ergebnisse des ersten Durchlaufs (Juni 2009) [43] sowie die Ergebnisse des jüngsten Durchlaufs (Juni 2011) einander gegenübergestellt. Die Anzahl der gefundenen Krankheiten konnte um ein Viertel (818) gesteigert werden, die Anzahl der Referenzen sogar um 32%. Die Anzahl der Kombination aus EC Nummer, Krankheit und der Referenz stieg in dieser Zeit um 15%.

Tabelle 3.12: Die Menge der identifizierten Kookkurrenzen beobachtet über die Zeit. In der Tabelle sind die Anzahl der Kombination aus EC Nummer, Krankheit und der Referenz, in der sie als kookkurrierende Entitäten vorliegen sowie deren jeweilige Anzahl aufgeführt, aufgeschlüsselt nach unterschiedlichen Referenzen, Krankheiten und Enzymen. Die Zahlen geben die Ergebnisse des ersten (Juli 2009) [43] und des aktuellsten (Juli 2011) Durchlaufs an sowie deren Anstieg absolut und prozentual.

Aktualisierungen	**Juli 2009**	**Juli 2011**	**Anstieg 2009-2011**	
Kombination	745.650	858.077	112.427	15%
Referenzen	395.776	522.096	126.320	32%
Krankheiten	3.259	4.077	818	25%
EC Nummern	1.863	2.062	199	11%
Präzision	0.840	0.839		
Vollständigkeit	0.882	0.894		
F_1 Maß	0.860	0.866		

Die Vollständigkeit ist um einen Prozentpunkt gestiegen. Die anderen Werte der Qualitätskenngrößen haben sich in diesem Zeitraum kaum verändert. Bei der Anpassungen und Optimierung der genutzten Wörterbücher werden bestimmte Synonyme aus den Wörterbüchern gestrichen, obwohl sie eine legitime Bezeichnung einer Entität darstellen. Diese Bezeichnungen sind jedoch gleichzeitig Homonyme, d.h. sie stellen ebenfalls die legitime Bezeichnung anderer Entitäten dar. Beispielsweise steht die englische Bezeichnung „strain“ für Muskelzerrung, die aber in der Systematik von Organismen den Begriff „Stamm“ bezeichnet. Dieser Begriff kommt 395.148 Mal in den PubMed (Stand Dezember 2009) Titeln und Kurzzusammenfassungen vor und wird weitaus häufiger in seiner systematischen Bedeutung als zur Beschreibung einer Muskelzerrung verwendet. Die Löschung solcher Begriffe vermindert zwar die Vollständigkeit und die Anzahl der unterschiedlichen Referenzen kann kurzfristig rückläufig sein, sorgt aber dafür, dass der falsch positive Ergebnisanteil sinkt. Es gibt Bezeichnungen und Akronyme, die zwar ein Homonym darstellen, deren Streichung aber nicht möglich ist, weil das Enzym oder die Krankheit bevorzugt mit diesem Begriff in wissenschaftlichen Artikeln benannt ist und der Anteil der falsch positiven Ergebnisse dafür in Kauf genommen werden muss, was gleichzeitig eine Minderung der Präzision bedingt. Das Akronym „ADH“ wird unter anderem für das Enzym Alkoholdehydrogenase (EC 1.1.1.1). In englischsprachiger, (bio-)medizinischer Literatur könnte es aber auch für „Adipic acid dihydrazide“, einer Chemikalie die im

Zusammenhang mit Steroid-Enzym-Immunoassays Verwendung findet, oder „Antidiuretic hormone“ stehen. Dennoch wird es so oft für die Bezeichnung der Enzymentität verwendet, dass es nicht gelöscht werden kann.

3.4. Die Klassifizierten von Entitätsbeziehungen

Durch die Kookkurrenzsuche konnte, wie in Abschnitt *3.3. Gemeinsam auftretende Krankheiten und Enzyme* gezeigt, eine Vielzahl von assoziierten Enzymen und Krankheiten gefunden werden. Um die gemeinsame Nennung von Enzymen und Krankheiten in einem Satz oder Titel auf die Art der semantischen Verbundenheit zu überprüfen, wurde eine angeschlossene Klassifizierung durchgeführt. Wenn eine semantische Verbundenheit vorlag, sollte diese Kategorien zugeordnet werden. Zu diesem Zweck wurden zuvor vier Kategorien definiert (*2.4.1. Definitionen der Entitätsrelationen* auf Seite 33): *causal interaction* (kausale Interaktion), *ongoing research* (Gegenstand der Erforschung), *diagnostic usage (diagnostische Nutzung)* und *therapeutic application* (therapeutische Anwendung).

Für die Umsetzung dieser Aufgabe wurde eine Methode des maschinellen Lernens angewendet, die auf der Klassifizierung durch eine Support Vector Machine (SVM) basiert. Zunächst wurde das entwickelte Verfahren getestet und optimiert. Die dabei erzielten Ergebnisse sind in Abschnitt *3.4.1. Test und Optimierung des Verfahrens* dargestellt. Die Ergebnisse der Klassifizierung und Einordnung der semantischen Verbindung von gemeinsam auftretenden Enzymen und Krankheiten in einem Satz oder Titel sind in Abschnitt *3.4.2. Anwendung auf den Klassifikationskorpus* beschrieben.

3.4.1. Test und Optimierung des Verfahrens

Einsatz von Stoppwörtern

Es wurde getestet, ob der Einsatz von Stoppwortlisten das Klassifizierungsergebnis verbessert. Vor der Berechnung der Termgewichte (*2.2.3. Aufbereitung und Termgewichtung* auf Seite 24) in der Vorverarbeitung wurden alle Wörter gelöscht, die auf der Stoppwortliste aufgeführt waren. Durch die Löschung von Stoppwörtern sollte zum einen eine Dimensionsreduktion der von der SVM zu verarbeitenden Vektoren erreicht werden, zum anderen eine Verstärkung der Termgewichte für Wörter mit höheren Informationsgehalt (siehe *Stoppwörter* in Abschnitt *3.1.1* auf Seite 48). Zum Vergleich wurde ein Klassifizierungsdurchlauf für die Kategorie *causal interaction* durchgeführt, bei dem gleichzeitig die auf unterschiedliche Weise vorverarbeiteten Eingaben durch die SVM klassifiziert wurden.

Die Vorverarbeitung wurde durchgeführt:

- ohne Löschung der Stoppwörter
- mit Löschung aller Stoppwörter auf der Liste *komplette Stoppwortliste*
- mit Löschung nur eines Teils der Stoppwörter *optimierte Stoppwortliste*

Eine Aufstellung der beiden alternativen Versionen der Stoppwortliste findet sich in Anhang B. *Liste der verwendeten statischen Stoppwörter* ab Seite 95.

In Tabelle 3.13 sind die maximal erreichten Werte der Präzision und der dazugehörigen Vollständigkeit für jede Form der Vorverarbeitung aufgeführt. Es wurden die maximalen Werte für die Präzision ausgewählt, die jeweils mit der dazugehörigen Vollständigkeit mindestens ein F_1 Maß von 0,6 erreichten.

Tabelle 3.13: Ein Vergleich der erreichten Präzision bei unterschiedlicher Stoppwortfilterung in der Vorverarbeitung. Angegeben ist die jeweils maximal erreichte Präzision und die dazu errechnete Vollständigkeit für die Klassifizierung ohne Stoppwortfilterung, Filterung mit einer optimierten Stoppwortliste und Filterung mit der vollständigen Stoppwortliste. Die Werte sind den Gesamtergebnissen entnommen unter der Voraussetzung, dass Präzision und Vollständigkeit zusammen mindestens eine F_1 Maß von 0,6 erreichten.

Ohne Löschung der Stoppwörter		**Mit Löschung der optimierten Stoppwörter**		**Mit Löschung aller Stoppwörter**	
Präzision	**Vollständigkeit**	**Präzision**	**Vollständigkeit**	**Präzision**	**Vollständigkeit**
0,741	0,545	0,721	0,577	0,705	0,536

Bei der Anwendung der kompletten Stoppwortliste wurden im Vergleich die schlechtesten Werte für die Präzision (0,705) und die Vollständigkeit (0,536) verzeichnet. Die Vorverarbeitung ohne Löschung der Stoppwörter erreichte zwar eine um 2% höhere Präzision im Vergleich zur Vorverarbeitung mit optimierter Stoppwortauswahl. Dem entgegen verzeichnet umgekehrt die Vorverarbeitung mit Löschung der optimierten Stoppwortauswahl eine um 3,2% höhere Vollständigkeit.

Es wurde auf eine Anwendung von Stoppwortlisten bei der Vorverarbeitung der Klassifizierungseingabedaten verzichtet, da die Kenngrößen der Qualitätsbewertung im Vergleich insgesamt nur geringe Abweichungen zeigten und der Zeitgewinn im Verarbeitungsschritt der Klassifizierung der SVM dadurch nicht signifikant verkürzt wurde.

Kreuzvalidierung

Bevor eine Verarbeitung des *Klassifikationskorpus* durchgeführt werden konnte, musste anhand von Sätzen und Titeln mit bekannter Einordnung in Kategorien ermittelt werden, ob und wie gut eine SVM die Klassifizierung für diese Aufgabe durchführen kann. Deshalb wurde zunächst eine fünffache Kreuzvalidierung durchgeführt, um die Effizienz weiterer Vorverarbeitungsvarianten zu testen und die Auswahl der optimalen Parametereinstellungen der SVM zu bestimmen. Dazu wurden die Sätze und Titel des Annotationskorpus mit mindestens einer Kookkurrenz von Enzym- und Krankheitsentitäten verwendet, mit und ohne Zuordnung zu den Kategorien (*3.2 Betrachtung des Annotationskorpus* auf Seite 52). Jedem Anteil an Sätzen und Titeln mit Zuordnung zu der jeweiligen Kategorie (Positivbeispiele) wurden die gleiche Menge an Sätzen und Titeln ohne Zuordnung (Negativbeispiele) beigefügt. Die Zusammenstellungen von Positiv- und Negativbeispielen wurden für jede Kategorie in fünf Teile gleicher Größe unterteilt. Reihum dienten vier Teile zum Trainieren der SVM und ein Teil zur Überprüfung der Klassifizierungsleistung, so dass jeder Teil einmal zur Überprüfung diente.

In der Kreuzvalidierung wurde eine Auswahl verschiedener Parametereinstellungen und aller Kernelfunktionen (linear, polynomial, radial Basis und sigmoidal), die in dem Programmpaket *SVMlight* zur Verfügung standen, getestet. Dies führte zu 2.688 unterschiedlichen Einstellungskonstellationen, aus denen jeweils ein Klassifikationsmodell errechnet wurde und die mit den Vorverarbeitungsvarianten *Löschung* und *Austausch* Arten (*2.2.3. Aufbereitung und Termgewichtung* auf Seite 24) getestet wurden.

Tabelle 3.14: Auflistung der maximal erreichten Werte des F_1 Maß in der fünffachen Kreuzvalidierung. In dieser Tabelle sind die Klassifizierungsmodelle mit den maximal erreichten Werten für das F_1 Maß sowie begleitend die entsprechenden Korrelationskoeffizienten nach Matthews und die Werte für die Fläche (AUC), der in Abbildung 3.2 dargestellten Kurven der receiver operating characteristics (ROC) aufgeführt. Die Auflistung ist aufgeschlüsselt nach den beiden unterschiedlichen Methoden der Vorverarbeitung Löschung (a) und Austausch (b) (2.2.3. Aufbereitung und Termgewichtung auf Seite 24).

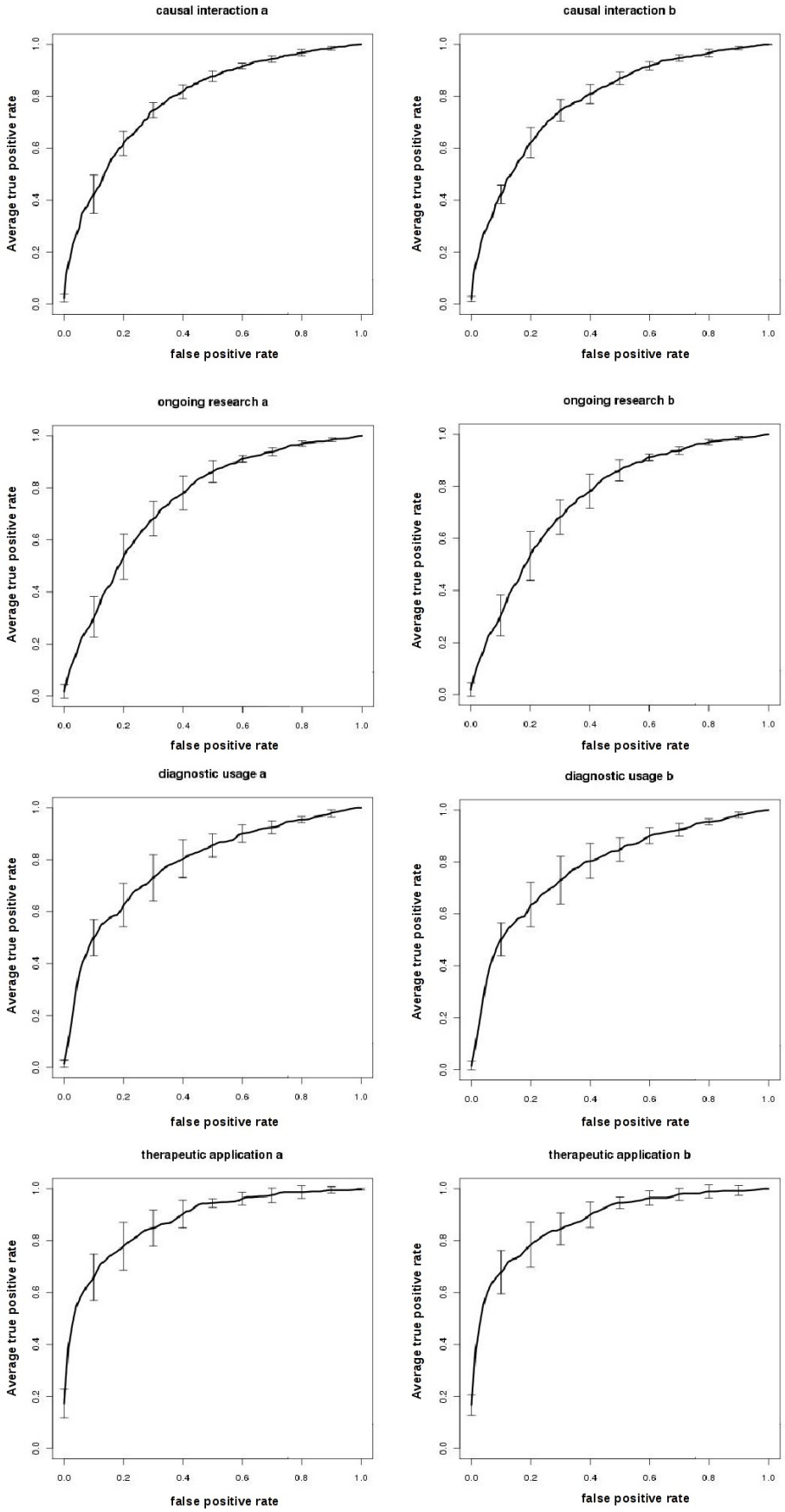
Kategorie	Vorverarbeitung	F_1 Maß	MCC	AUC
therapeutic application	a	0,802 ± 0,032	0,583 ± 0,076	0,878 ± 0,040
ongoing research	a	0,733 ± 0,024	0,395 ± 0,058	0,752 ± 0,037
diagnostic usage	a	0,738 ± 0,032	0,430 ± 0,074	0,784 ± 0,032
causal interaction	a	0,743 ± 0,009	0,429 ± 0,023	0,788 ± 0,020
therapeutic application	b	0,792 ± 0,016	0,548 ± 0,041	0,878 ± 0,038
ongoing research	b	0,744 ± 0,020	0,427 ± 0,051	0,752 ± 0,037
diagnostic usage	b	0,732 ± 0,022	0,412 ± 0,075	0,783 ± 0,033
causal interaction	b	0,742 ± 0,010	0,428 ± 0,024	0,786 ± 0,020

In Tabelle 3.14 sind die maximal erreichten Werte für das F1 Maß sowie begleitend die entsprechenden Korrelationskoeffizienten nach Matthews und die Werte für die Fläche (AUC) unter der in Abbildung 3.2 dargestellten Kurven der receiver operating characteristics (ROC) aufgeführt. Das F_1 Maß variiert zwischen 0.802±0.032 und 0.738±0.033 (0.583±0.076 und 0.395±0.058 entsprechender MCC). Ein Resultat der Kreuzvalidierung war der eher marginale Unterschied zwischen den Vorverarbeitungsmethoden im Hinblick auf die Werte für F_1 Maß und area under the curve (AUC). Die Vorverarbeitungsmethoden bestanden jeweils aus der *Löschung* (a) oder dem *Austausch* (b) der Namen, der in den Sätzen und Titeln enthaltenen Krankheits- und Enzymentitäten. Nur der Korrelationskoeffizient nach Matthews (MCC) unterscheidet sich bei den Kategorien mit geringerer oder mittlerer Anzahl von Positiv- und Negativbeispielen (*therapeutic application*, *ongoing research* und *diagnostic usage*). Der MCC, der beiden betrachteten Klassifizierungsmodelle der Kategorie *causal interaction* in Tabelle 3.14, differierte nur um 0,001 und zeigte somit keinen signifikanten Unterschied zwischen den beiden Methoden der Vorverarbeitung und wies die geringste Standardabweichung aller aufgeführten MCC Werte auf. Die Kategorie mit dem besten Wert für das F_1 Maß war *therapeutic application* mit 0,802±0,032 für die Vorverarbeitung durch *Löschung* und 0,792±0,016 für die Vorverarbeitung durch *Austausch*. Für diese Kategorien konnten auch die maximalen Werte für AUC (0,878±0,040 *Löschung*, 0,878±0,038 *Austausch*) und MCC (0,583±0,076 *Löschung*, 0,548±0,041 *Austausch*) verzeichnet werden. Die Werte für das F_1 Maß der Kategorien *causal interaction* und *ongoing research* sind ähnlich.

Jedoch ist die Vorverarbeitungsweise *Austausch*, mit einen Wert für das F_1 Maß von 0,744±0,020 und für den MCC von 0,427±0,051 erfolgreicher in der Kategorie *ongoing research* als die *Löschung* mit einen Wert für das F_1 Maß von 0,733±0,024 bzw. einem MCC von 0,395±0,058. In allen anderen Kategorien zeigt die Vorverarbeitung durch *Löschung* der Namen der Enzym- und Krankheitsentitäten gleiche oder geringfügig bessere Werte.

Die in Abbildung 3.2 dargestellten ROC Kurven entsprechen den Klassifikationsmodellen, deren weitere ermittelte Werte (F_1 Maß, MCC, AUC) in Tabelle 3.14 aufgeführt sind. Jeweils übereinander sind die ROC Graphen für die Kategorien *causal interaction*, *ongoing research*, *diagnostic usage* und *therapeutic application* und jeweils nebeneinander sind die Methoden der Vorverarbeitung, *Löschung* (a) und *Austausch* (b) aufgeführt. Es handelt sich dabei um vertikal gemittelte (gemittelte true positive rate) ROC Graphen aus den fünf Durchläufen für die jeweiligen Klassifizierungsmodelle in der Kreuzvalidierung. Obwohl die Kategorie *therapeutic application* die wenigsten Trainingsbeispiele umfasst, zeigten die ROC Kurven der Klassifizierungsmodelle (Abbildung 3.2, *therapeutic application* a und b) die größten Flächen (Tabelle 3.14, AUC). Die Kurven mit der geringsten Standardabweichung waren die der Kategorie *causal interaction* (Abbildung 3.2, *causal interaction* a und b), die zugleich die Kategorie mit den meisten Trainingsbeispielen im *Annotationskorpus* gewesen ist.

Abbildung 3.2: Receiver operating characteristic (ROC) Kurven, der Klassifizierungsmodelle, die maximale Werte der F_1 Maße erreichten.



Eignung der Methode zur Klassifizierung

Anhand der Ergebnisse der fünffachen Kreuzvalidierung konnte gezeigt werden, dass die Klassifizierung semantischer Relationen mit der gewählten Methode möglich ist. Die angewendeten Vorverarbeitungsschritte umfassen keine Methoden der natürlichen Sprachauswertung (natural language processing), wie Lemmatisierung oder der Zuordnung von Wörtern zu den entsprechenden Wortarten (part-of-speech tagging), sondern nur eine sehr eingeschränkte Form der linguistischen Vorverarbeitung, wie das Löschen oder der Austausch von Entitätsnamen und der anschließenden Berechnung von Termgewichten. Es konnte bestätigt werden, dass die Qualität der Klassifizierung nicht eingeschränkt ist und die Effektivität der Berechnungen durch die SVM nicht durch die hohe Dimensionalität des Vektorraums eingeschränkt wird [49]. Sowohl das Löschen als auch der Austausch von Namen von Krankheits- und Enzymentitäten brachten vergleichbar gute Resultate. Je nach Kategorie konnte eine leichte Verbesserung der Leistung entweder durch *Löschung* oder *Austausch* beobachtet werden. Bei der Anwendung der Methode auf den Klassifikationskorpus wurden deshalb beide Methoden beibehalten.

Signaldeutlichkeit

Die Kategorie *therapeutic application* schien die deutlichsten Unterscheidungssignale zu beinhalten, obwohl sie die wenigsten Trainingsbeispiele im *Annotationskorpus* umfasst. Bei der Bestimmung der Urteilerübereinstimmung, anhand des κ-Koeffizienten (siehe *Diversität der Interpretation einer Relation* in Abschnitt *3.2.* auf Seite *54*), war *therapeutic application* ebenfalls die Kategorie mit dem höchsten Wert an Übereinstimmung zwischen den Annotatoren. Es ist offensichtlich, dass dort, wo der Mensch eine eindeutige Zuordnung vornimmt, auch der Algorithmus der SVM effizienter arbeitet.

Die Wahl eines Schwellenwertes

Die Auftragungen als ROC Kurven (Abbildung 3.2) zeigen, wie eine Variation des Schwellenwertes die Größe der richtig positiven und richtig negativen Ergebnisanteile verändert. Viele Fragestellungen einer Klassifizierung bedingen, dass der zu erwartende Anteil der real positiven Instanzen im Vergleich zu den negativ Instanzen eher gering ist. Deswegen ist das Klassifikationsmodell sowie der Schwellenwert für ein kostenoptimiertes Klassifikationsverhalten (*2.5.4 Graphische Validierung von Klassifikatoren* ab Seite 42) unter diesen Bedingungen zu bevorzugen [73]. Das

bedeutet, dass für die Klassifizierung des *Klassifikationskorpus* Schwellenwerte gewählt werden, die nur eine positive Klassifizierung erlauben, sofern es eine hoch gewichtete Evidenz dafür gibt (hohe Präzision/niedrige Vollständigkeit). Um sukzessive die Vollständigkeit zu erhöhen, kann der Schwellenwert abgesenkt werden (geringere Präzision/erhöhte Vollständigkeit). Um sowohl Klassifizierungsergebnisse mit hoher Präzision als auch hoher Vollständigkeit zu erzielen, wurde deshalb bei der Anwendung auf den Klassifikationskorpus (siehe folgender Abschnitt 3.4.2.) ein System mit Qualitätsstufen von (1 bis 4) etabliert, dass dieser Abstufung Rechnung trägt.

3.4.2. Anwendung auf den Klassifikationskorpus

Der *Klassifikationskorpus* wurde aus den 624.836 Sätzen von Kurzzusammenfassungen und 187.601 Titeln gebildet, in denen Kookkurrenzen von Enzym- und Krankheitsentitäten ermittelt wurden (*3.3. Gemeinsam auftretende Krankheiten und Enzyme*). Die Sätze und Titel in denen Enzyme und Krankheiten gemeinsam auftreten, sollten durch die Klassifizierung mit einer SVM bei entsprechend vorhandenem semantischen Kontext den definierten Kategorien *causal interaction* (kausale Interaktion), *ongoing research* (Gegenstand der Erforschung), *diagnostic usage* (diagnostische Nutzung) und *therapeutic application* (therapeutische Anwendung) zugeordnet werden (*2.4.1. Definitionen der Entitätsrelationen* ab Seite 33).

Für Training und Test der Klassifizierung durch die SVM wurden die Sätze und Titel des *Annotationskorpus* mit mindestens einer Kookkurrenz von Enzym- und Krankheitsentitäten verwendet, mit und ohne Zuordnung zu den Kategorien (*3.2 Betrachtung des Annotationskorpus* ab Seite 52). Jedem Anteil an Sätzen und Titeln mit Zuordnung zu der jeweiligen Kategorie (Positivbeispiele) wurden die gleiche Menge Sätze und Titel ohne Zuordnung (Negativbeispiele) beigefügt. Die Zusammenstellungen von Positiv- und Negativbeispielen wurden zum Trainieren und zur Qualitätsprüfung, für jede Kategorie in fünf Teile gleicher Größe unterteilt. Zum Trainieren der SVM, also der Erstellung der Klassifizierungsmodelle mit den in der Kreuzvalidierung ermittelten optimalen Einstellungen, dienten vier Teile. Der Kookkurrenzkorpus wurde zusammen mit dem verbliebenen Teil durch die SVM klassifiziert. Dieser fünfte Teil diente zur anschließenden Qualitätsüberprüfung anhand verschiedener Kenngrößen (*2.5.2. Skalare Kenngrößen* ab Seite 39).

In jeder Kategorie wurden die Klassifizierungsergebnisse in vier Qualitätsstufen eingeteilt, wobei sich *Qualitätsstufe 4* durch die höchste Präzision und Spezifität und *Qualitätsstufe 1* durch die höchste Vollständigkeit auszeichnen. Die Übergänge von *Qualitätsstufe 1* zu *Qualitätsstufe 4* begleiten eine Erhöhung der Präzision und Spezifität, welches ein Absinken der Vollständigkeit bedingt.

In den Tabellen 3.15 bis 3.18 sind die Ergebnisanteile der Klassifizierung für jede Kategorie in den jeweiligen Qualitätsstufen dargestellt. Es sind jeweils die Anzahl der Kombinationen aus EC Nummern, Krankheit und PubMed Referenz und die dazu ermittelten Werte der Qualitätskenngrößen angegeben. In jeder Kategorie konnte in der Qualitätstufe 4 eine Spezifität von mindestens 96,7% und eine Präzision zwischen 97,6% und 85,5% erreicht werden.

Tabelle 3.15: Klassifizierungsergebnisse der Kategorie causal interaction

Qualitätsstufe	Unterschiedliche Kombinationen von EC nummern, Krankheiten und Referenzen	Präzision	Vollständigkeit	Genauigkeit	Spezifität	MCC
4	200.137	0,855	0,192	0,580	0,967	0,252
3	473.517	0,778	0,533	0,690	0,848	0,401
2	480.782	0,779	0,536	0,692	0,848	0,404
1	648.545	0,702	0,775	0,723	0,670	0,448

Tabelle 3.16: Klassifizierungsergebnisse der Kategorie ongoing research

Qualitätsstufe	Unterschiedliche Kombinationen von EC nummern, Krankheiten und Referenzen	Präzision	Vollständigkeit	Genauigkeit	Spezifität	MCC
4	68.596	0,889	0,137	0,560	0,983	0,225
3	313.954	0,744	0,547	0,680	0,812	0,372
2	356.341	0,740	0,607	0,697	0,786	0,400
1	580.170	0,642	0,812	0,680	0,547	0,372

Tabelle 3.17: Klassifizierungsergebnisse der Kategorie therapeutic application

Qualitätsstufe	Unterschiedliche Kombinationen von EC nummern, Krankheiten und Referenzen	Präzision	Vollständigkeit	Genauigkeit	Spezifität	MCC
4	158.143	0,976	0,548	0,767	0,986	0,594
3	182.532	0,938	0,616	0,788	0,959	0,612
2	345.421	0,855	0,808	0,836	0,863	0,672
1	422.601	0,831	0,877	0,849	0,822	0,700

Tabelle 3.18: Klassifizierungsergebnisse der Kategorie diagnostic usage

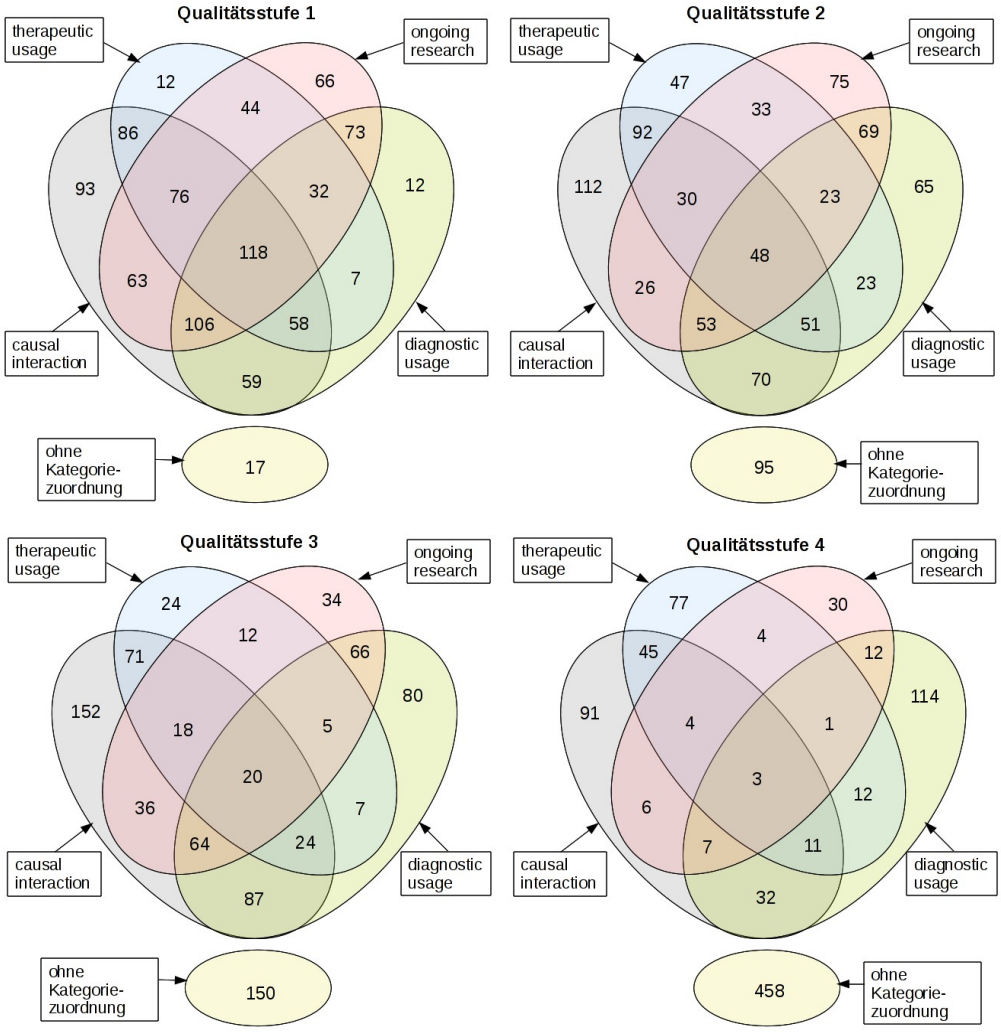
Qualitätsstufe	Unterschiedliche Kombinationen von EC nummern, Krankheiten und Referenzen	Präzision	Vollständigkeit	Genauigkeit	Spezifität	MCC
4	193.632	0,939	0,326	0,653	0,979	0,403
3	354.083	0,833	0,632	0,753	0,874	0,521
2	401.714	0,795	0,695	0,758	0,821	0,520
1	454.540	0,764	0,716	0,747	0,779	0,496

Für die Kategorie *therapeutic application* wurde in der Qualitätsstufe 4 die beste Präzision (0.976) in Kombination mit der Volltändigkeit (0,548) (Tabelle 3.17) erreicht. Im Hinblick auf die Anzahl der Ergebnisse ist die Kategorie *causal interaction* in allen Qualitätsstufen führend (Tabelle 3.15) im Vergleich zu den übrigen Kategorien. Da eine Kookkurrenz auch mehreren Kategorien zugeordnet werden kann, können sich Überschneidungen der Ergebnisanteile ergeben. Diese werden im folgenden Abschnitt *Überschneidungen der Kategorien* näher betrachtet.

Überschneidungen der Kategorien

Sofern mehrere Aussagen innerhalb eines Satzes oder Titels enthalten sind, in dem kookkurrierende Enzym- und Krankheitsentitäten auftreten, kann es erforderlich sein eine Einordnung in mehr als eine Kategorie vorzunehmen. Ebenso ist es möglich, dass eine Kurzzusammenfassung mehrere Sätze mit Kookkurrenzen enthält, die jede für sich einer anderen Kategorie von semantischer Beziehung entspricht.

Abbildung 3.3: Die Verteilung der Schnittmengen der Kategorien. Angegeben ist die Größe der Schnitt- und Differenzmengen der Kombinationen (Zahlen x 10^3, gerundet) von EC Nummern, Krankheiten und Referenzen, die jeweils einer oder mehreren Kategorien causal interaction (grau), therapeutic application (blau), ongoing research (rosa) und diagnostic usage (grün) in den entsprechenden Qualitätsstufe 1 bis 4 durch die Klassifizierung zugeordnet wurden. Die Anzahl der Kombinationen die keiner Kategorie zugeordnet wurden, ist in den Abbildungen für jede Qualitätsstufe jeweils unterhalb in dem gelben Oval angegeben.



In Abbildung 3.3 ist die Verteilung der Schnitt- und Differenzmengen der Kombinationen aus EC Nummer, Krankheit und PubMed Referenz der jeweiligen Kategorien für jede Qualitätsstufe dargestellt. Große Teilmengen der Kategorien *therapeutic application*, *diagnostic usage* und *ongoing research* zeigten Überschneidungen mit den Kombinationen, die ebenfalls der Kategorie *causal interaction* zugeordnet wurden. In jeder Qualitätsstufe waren durchschnittlich 50% der Kombinationen, die den jeweiligen Kategorien *therapeutic application*, *diagnostic usage* oder *ongoing research* zugeordnet worden sind, auch Teil der Kategorie *causal interaction* (Abbildung 3.3).

Enzyme in multipler Beziehung zu einer Krankheit

Erklären lassen sich die Überschneidungen mit der Kategorie *causal interaction* (Abbildung 3.3) dadurch, dass die Veränderung des Enzyms das eine Krankheit auslösen kann auch im Zusammenhang mit der Diagnose oder der Therapie der Krankheit wichtig werden kann. Bei Morbus Gaucher ist eine Mutation im Gen des Enzyms Glukozerebrosidase (EC 3.2.1.45) dafür verantwortlich, dass der enzymatische Abbau von Membranbestandteilen (Glukosylzeramid) abgestorbener Zellen nicht vollständig erfolgt und es zu einer Anreicherung in Makrophagen des Blutsystems und in Histiozyten verschiedener Organgewebe kommt, die wiederum u.a. schwere neurologische Ausfallserscheinungen, Störungen der Organfunktion und Anämien auslösen kann [88]. Die Ursache der Krankheit wird durch eine lebenslange Enzymersatztherapie behandelt [89]. So ist hier das Enzym gleichzeitig das therapeutische Mittel.

In den höheren Qualitätsstufen nimmt der Anteil der Überschneidungen ab und der Anteil der Kombinationen, die keiner Kategorie zugeordnet sind, steigt an. Das lässt darauf schließen, dass nur Aussagen die eindeutig sind den Kategorien zugeordnet werden und Sätze mit schwächeren oder uneindeutigen Aussagen zurückgewiesen werden.

Die Prostaglandin Endoperoxidsynthase

Prostaglandin Endoperoxidsynthase (EC 1.14.99.1) katalysiert einen initialen Schritt der Prostaglandinsynthese und ist somit ein Enzym mit zentraler Bedeutung für den Biosyntheseweg von der Arachidonsäure zu den Prostaglandinen. Die Verteilung des Enzyms im Gewebe und seine physiologische Funktionen im Einzelnen sind abhängig von seinen Isoformen. Das Enzym ist beteiligt an der Aggregation von Thrombozyten, dem Schutz der Magenschleimhaut und der renalen Elektrolythomöostase [90]. Die

induzierbare Isoform spielt eine Rolle bei Entzündungsprozessen und seine Expression, die Biosynthese aus genetischer Information, wird ausgelöst durch Stimuli wie Zytokine, Wachstumsfaktoren oder Hormone. Die Überexpression des Isoenzyms beobachtet man bei präkanzerösen Vorstufen von Gewebeveränderungen und bereits malignen (bösartigen) Tumoren von Darm, Leber, Pankreas, Brust, Lunge, Blase, Haut, Magen, Hals und Nacken [91]. Die Prostaglandin Endoperoxidsynthase ist das Wirkziel von nichtsteroidalen Entzündungshemmern wie Ibuprofen und Acetylsalicylsäure und somit relevant bei der Behandlung von leichten bis mäßigen Schmerzen, Fieber und Entzündungsprozessen sowie bei kardiovaskulären und rheumatischen Erkrankungen.

Prostaglandin Endoperoxidsynthase ist eines der Enzyme, die mit am häufigsten (in 9.377 Referenzen) zusammen mit einer Krankheit gefunden wurde (*3.3. Gemeinsam auftretende Krankheiten und Enzyme*). Das Enzym ist in 15.103 Kombinationen aus EC Nummer, Krankheit und PubMed Referenz vertreten und darin mit 832 verschiedenen Krankheiten assoziiert. Es befindet sich in jeder Schnittmenge aller vier Kategorien (Abbildung 3.3) und in jeder Qualitätsstufe unter den zehn häufigsten Enzymen. Durch seine vielfache Bedeutung in Physiologie und Pathophysiologie sind diese Häufungen in den Ergebnissen plausibel und spiegeln die real existierenden Verbindungen mit den unterschiedlichsten Krankheiten wieder.

3.4.3. Vergleich der Kategorie *therapeutic application* mit DrugBank

Die DrugBank [92] Datenbank der University of Alberta enthält mehr als 4.700 Einträge zu Arzneimitteln und bietet ausführliche Informationen über Medikamenteninhaltsstoffe und ihre Wirkziele. Es wurde ein Vergleich mit den Einträgen der DrugBank (Version 2.5) vorgenommen, um die spezifische Relevanz der Einträge in der Kategorie *therapeutic application* anhand einer externen Datenquelle zu überprüfen. Die 39.023 Einträge zu Wirkzielen in DrugBank umfassen 9.713 unterschiedliche Wirkzielbezeichnungen. Von den Wirkzielbezeichnungen konnten 1.187 als Bezeichnung für eine Enzymentität identifiziert und so einer EC Nummer zugeordnet werden. Aus den Freitexteinträgen zur Indikation für die therapeutische Anwendung in DrugBank konnten 264 Krankheitsentitäten identifiziert werden. Diese Anzahl repräsentiert einen Anteil von 7% der unterschiedlichen Krankheiten, die in den Ergebnissen der Kategorie *therapeutic application* enthalten sind. Ein Vergleich der EC Nummern der Einträge von DrugBank und der EC Nummern in der Kategorie *therapeutic application* ohne Berücksichtigung der Krankheit zeigte, dass ein Anteil zwischen 72,8% und 76,4% (in Abhängigkeit der Qualitätsstufe) der Kombinationen aus EC Nummer, Krankheit und PubMed über die EC Nummer bestätigt werden konnte

(Tabelle 3.19). Der Anteil an Übereinstimmungen, unter zusätzlicher Berücksichtigung der Krankheiten, ist entsprechend der geringen Anzahl identifizierter Krankheiten in DrugBank geringer.

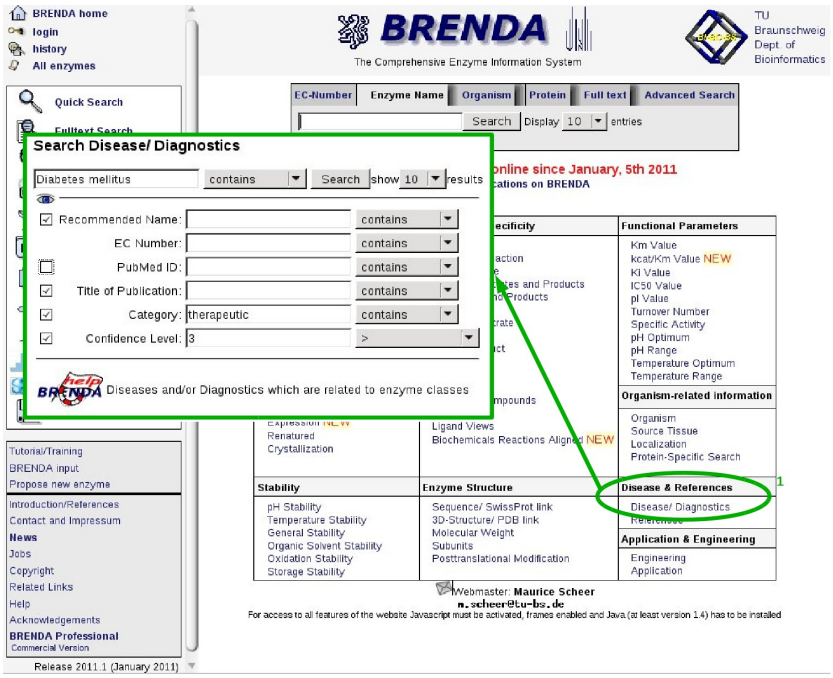
Tabelle 3.19: Vergleich der Einträge der DrugBank Datenbank mit den Ergebnisanteilen, die bei der Klassifizierung der Kategorie therapeutic application zugeordnet wurden. Aufgeführt sind die unterschiedlichen EC Nummern, die in der Kategorie therapeutic application enthalten sind und der absolute und prozentuale Anteil, der in Übereinstimmung mit DrugBank bestätigt werden konnte. Daneben ist angegeben wie hoch der übereinstimmende Anteil der Kombinationen aus EC Nummer, Krankheit und PubMed Referenz ist, wenn EC Nummern bzw. EC Nummern und Krankheiten mit DrugBank verglichen werden.

Qualitätsstufe	EC Nummern		Kombinationen EC Nummer, Krankheit und PubMed Referenz		
	enthalten in *therapeutic application*	Übereinstimmung *therapeutic application* / DrugBank	*therapeutic application*	Übereinstimmung *therapeutic application* / DrugBank	Übereinstimmung *therapeutic application* / DrugBank mit Berücksichtigung von Krankheiten
1	1.622	863 (53,2%)	422.601	307.477 (72,8%)	18.834 (4,5%)
2	1.516	831 (54,8%)	345.421	251.460 (72,8%)	17.018 (4,9%)
3	1.227	710 (57,9%)	182.532	137.701 (75,4%)	12.466 (6,8%)
4	1.183	694 (58,7%)	158.143	120.760 (76,4%)	11.448 (7,2%)

3.4.4. Integration in die BRENDA Informationsplattform

Die Ergebnisse der Kookkurrenzsuche von Krankheits- und Enzymentitäten in Kombination mit der Klassifizierung der semantischen Beziehungen sind in die Internetpräsenz des BRENDA Informationssystem (http://www.brenda-enzymes.org) integriert worden [93]. Die Ergebnisse der automatischen Klassifizierung sind seit der BRENDA Aktualisierung 2011/1 integriert. Die Einträge zu Krankheiten, die in einer Verbindung zu Enzymen stehen, können dort in der Sektion “Disease/Diagnostics” (Abbildung 3.4) abgerufen werden und sind anhand einer speziellen Abfragemaske durchsuchbar.

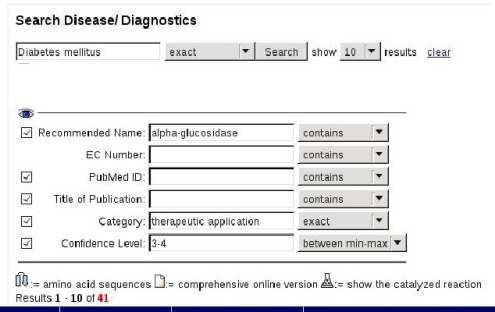
Abbildung 3.4: Das BRENDA Internetportal. Durch die Auswahl „Disease/Diagnostics“ gelangt man zu dem entsprechenden Suchformular (grün umrandet, vergrößert dargestellt).



Durch das Suchformular (Abbildung 3.5, a) können Abfragen der Ergebnisse spezifiziert werden. Die Suchfelder erlauben eine Suche nach einem Teil oder der vollen Bezeichnung einer Krankheit, eines Enzyms, einer EC Nummer oder dem Titel der Publikation. Diese Felder können individuell kombiniert oder separat genutzt werden, um das Suchmuster der Abfrage optimal zu verfeinern. Durch die Auswahl des Markierungsfelds “Category” oder “Confidence Level” werden die Kategorien sichtbar, denen die Referenz durch die Klassifizierung zugeordnet worden ist. Nach der Übermittlung der Abfrage wird eine Ergebnistabelle mit den jeweils gewählten Feldern erstellt (Abbildung 3.5, b). Jeder Eintrag enthält eine verlinkte Anbindung zur BRENDA „flat file“ Ansicht des entsprechenden Enzyms, zu einer Suche nach Aminosäuresequenzen des Enzyms, einer Ansicht der vom Enzym katalysierten Reaktionen und zur Kurzzusammenfassung der Referenz in der PubMed Datenbank. Auf diese Weise wurden die krankheitsbezogenen Einträge voll in das BRENDA Informationssystem integriert.

Abbildung 3.5: Der Zugang zu krankheitsbezogenen Informationen in BRENDA. Die Abfragemaske (a) enthält verschiedene Felder um ein abgestimmtes Suchmuster durch Eingabe zusammenzustellen. Die Felder können beliebig kombiniert werden, um die Suche zu verfeinern und die Ergebnismenge einzugrenzen. Exemplarisch dargestellt ist eine Ergebnistabelle (b) für die Suchanfrage (a) nach "Diabetes mellitus" in Verbindung mit der Enzymbezeichnung „alpha glucosidase" und einer Zuordnung in die Kategorie „therapeutic application" in den Qualitätsstufen 3 und 4.

a



b

EC Number	Recommended Name	Disease	PubMed ID	Title of Publication	Category	Confidence Level
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	1490691	alpha-Glucosidase inhibition in the treatment of diabetes mellitus.	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	8001622	An evaluation of the potential side-effects of alpha-glucosidase inhibitors used for the management of diabetes mellitus.	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	8562269	[alpha-Glucosidase inhibitors in the therapy of diabetes mellitus]	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	9430016	Safe and effective treatment of diabetes mellitus associated with chronic liver diseases with an alpha-glucosidase inhibitor, acarbose.	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	9589946	[The use of the alpha-glucosidase inhibitor acarbose for the treatment of type-2 diabetes mellitus in secondary sulfanilamide resistance]	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	9702468	Long-term effectiveness of a new alpha-glucosidase inhibitor (BAY m1099-miglitol) in insulin-treated type 2 diabetes mellitus.	therapeutic application	4
3.2.1.20	alpha-glucosidase	Diabetes Mellitus	9739502	Potential of alpha-glucosidase inhibitors in elderly patients with diabetes mellitus and impaired glucose tolerance.	therapeutic application	4

Durch diese Integration der Ergebnisse in das BRENDA Informationssystem ist es möglich gezielt nach Referenzen zu suchen, die relevante Informationen zu Enzymen und Krankheiten enthalten. Durch die Eingrenzung der Kategorie kann man die Zahl der gefunden Referenzen beschränken und erhält nur Referenzen von denen angenommen wird, dass Informationen zu Enzymen und Krankheiten enthalten sind, die dem Spektrum der definierten Kategorien *causal interaction* (kausale Interaktion), *ongoing research* (Gegenstand der Erforschung), *diagnostic usage* (diagnostische Nutzung) und *therapeutic application* (therapeutische Anwendung) entsprechen. Durch die Wahl der höheren Qualitätsstufen lässt sich die Referenzanzahl weiter eingrenzen auf Referenzen deren Einordnung in dieser Kategorie einer jeweils höheren Evidenz zugrunde liegt.

3.5. Zugangsnummern biologischer Datenbanken

Für die Suche nach Zugangsnummern biologischer Datenbanken wurde der *Volltextkorpus* verwendet. Im Folgenden werden die Ergebnisse der regelbasierten Suche nach Zugangsnummern aufgeführt und diskutiert.

Die regelbasierte Suche

Insgesamt konnten in 29.927 Referenzen 112.522 unterschiedliche Zugangsnummern identifiziert werden. In Tabelle 3.20 ist die Anzahl der identifizierten Zugangsnummern nach ihren jeweiligen Ursprungsdatenbanken aufgeschlüsselt. Es wurde auch nach Wörtern oder Wortfolgen gesucht, die die korrekte Einordnung als Zugangsnummer sowie die Zuordnung zur entsprechenden Datenbank bestätigen sollten (Indikatorterm) und deren Zeichenabstand zu der Zugangsnummer festgehalten.

Tabelle 3.20: Die Anzahl der extrahierte Zugangsnummern aufgeschlüsselt nach Ursprungsdatenbank.

Datenbank	Anzahl Zugangsnummern	Zugangsnummern mit einem begleitenden Indikatorterm	Zugangsnummern mit einem begleitenden Indikatorterm Zeichenabstand < 100
GenBank	43.859	29.200	15.101
PDB	18.131	18.131	4.809
NCBI RefSeq	15.744	9.085	5.684
UniProt/SwissProt/TREMBL	13.677	8.162	4.743
EMBL	12.640	8.074	4.230
DDBJ	8.253	5.495	2.608
PROSITE	151	136	53
NCBI	67	31	16
∑ alle Datenbanken	112.522	78.314	37.244

Der Anteil der extrahierten GenBank Zugangsnummern beträgt 40% an der Gesamtmenge aller extrahierten Zugangsnummern aller Datenbanken und ist der größte Anteil im Vergleich mit den Ergebnisanteilen für die übrigen Datenbanken. Der Anteil der extrahierten UniProt/SwissProt/TREMBL Zugangsnummern (im weiteren kurz UniProt genannt) beträgt 12%. In der Ergebnismenge wurden keine extrahierten PDB Zugangsnummern zugelassen, die keinen begleitenden Indikatorterm aufwiesen, weil das Format, der im Vergleich kurzen vierstelligen Zugangsnummer von PDB einen sehr hohen Anteil falsch positiver Einträge ermöglicht. Deshalb ist der Anteil der insgesamt extrahierten Zugangsnummern und der Anteil der Zugangsnummern mit einem

begleitenden Indikatorterm identisch. Selbst mit dieser strikten Einschränkung ist der Anteil der extrahierten PDB Zugangsnummern mit 16% der zweitgrößte, nach dem Anteil der UniProt Zugangsnummern.

Anhand des Formats der Zugangsnummer wurde die Art des jeweils referenzierten Eintrags bestimmt. In Tabelle 3.21 ist die jeweilige Anzahl der extrahierten Zugangsnummern, aufgeschlüsselt nach der Art des Eintrags sowie der Datenbank, aufgeführt. Diese Aufschlüsselung erfolgte durch die abgeleiteten Formatierungsregeln der Zugangsnummern. Eine ausführliche Auflistung der Formatierungsregeln befindet sich im Anhang in Tabelle Anhang 5 und Tabelle Anhang 6.

Tabelle 3.21: Die Anzahl der extrahierte Zugangsnummern aufgeschlüsselt nach Art des durch die Zugangsnummer referenzierten Eintrags. Die Aufschlüsselung erfolgt nach den jeweiligen Datenbanken sowie den durch die Zugangsnummern referenzierten Sequenztypen.

Datenbank	Art des Eintrags	Anzahl Zugangsnummern
UniProt/SwissProt/TREMBL	Aminosäuresequenz	13.677
GenBank	Aminosäuresequenz	10.926
NCBI RefSeq	Aminosäuresequenz	9.495
EMBL	Aminosäuresequenz	3.563
∑ alle Datenbanken	Aminosäuresequenz	37.661
GenBank	Nucleotidsequenz	32.490
EMBL	Nucleotidsequenz	8.750
DDBJ	Nucleotidsequenz	8.069
NCBI RefSeq	Nucleotidsequenz	790
NCBI	Nucleotidsequenz	67
∑ alle Datenbanken	Nucleotidsequenz	50.166
GenBank	Whole-Genome-Shotgun Sequenzen	443
EMBL	Whole-Genome-Shotgun Sequenzen	327
DDBJ	Whole-Genome-Shotgun Sequenzen	183
NCBI RefSeq	Whole-Genome-Shotgun Sequenzen	149
∑ alle Datenbanken	Whole-Genome-Shotgun Sequenzen	1.102
PDB	Protein Struktur	18.131
NCBI RefSeq	mRNA-Sequenz	5.310
PROSITE	Protein Domäne /Familie/Gruppe	151
DDBJ	Mass sequences for Genome Annotation	1

Aus Tabelle 3.21 ist ersichtlich, dass 44,6% der extrahierten Zugangsnummern Einträge für Nucleotidsequenzen und 33,5% der extrahierten Zugangsnummern Einträge für Aminosäuresequenzen in den jeweiligen Datenbanken referenzieren. Die Einträge *Whole-Genome-Shotgun Sequenzen* und *Mass sequences for Genome Annotation* stehen ebenfalls für Nucleotidsequenzen. In den Ursprungsdatenbanken wird jedoch bei der Vergabe einer Zugangsnummer nach dieser Art der Sequenzierung unterschieden und diese Unterscheidung durch ein anderes Zugangsnummernformat kenntlich gemacht.

Vergleich mit den UniProt Zugangsnummern

Um zum einen den Anteil der gültigen Zugangsnummern zu bestimmen und zum anderen den Einfluss einer Eingrenzung durch Filterung von Ergebnissen ohne begleitenden Indikatorterm zu bestimmen, wurden die extrahierten Zugangsnummern, die der Datenbank UniProt zugeordnet wurden, mit in der UniProt Datenbank eingetragenen Zugangsnummern verglichen. In Tabelle 3.22 sind die Ergebnisse für den Vergleich der bekannten Zugangsnummern der UniProt Datenbank (Stand Dezember 2010) mit den extrahierten UniProt Zugangsnummern aufgeführt.

Tabelle 3.22: Ein Vergleich der in UniProt Datenbank (Stand Dezember 2010) eingetragenen UniProt (gültigen) Zugangsnummern mit den UniProt zugeordneten Zugangsnummern, aus den Ergebnissen der regelbasierten Suche. Angegeben sind jeweils die Anzahl der übereinstimmenden Zugangsnummern im Vergleich, der prozentuale Anteil der Übereinstimmungen an der Gesamtmenge der extrahierten Zugangsnummern für UniProt sowie der prozentuale Anteil der Übereinstimmungen an der Gesamtmenge der extrahierten Zugangsnummern für UniProt unter den beiden Bedingungen: 1. eines vorhandenen Indikatorterms (Anteil in Prozent gesamt); 2. eines Indikatorterms im Zeichenabstand von maximal 100 Zeichen zu der extrahierten Zugangsnummer (Anteil in Prozent eingeschränkt).

Art des Vergleichs	Anzahl	Anteil in Prozent gesamt (Anteil in Prozent eingeschränkt)
Anzahl gültige UniProt Zugangsnummern	13.086.732	-
Anzahl extrahierte UniProt Zugangsnummern	13.677	100%
Übereinstimmungen gültige UniProt / extrahierte UniProt Zugangsnummern	12.921	94,5% (-)
Anzahl extrahierte UniProt Zugangsnummern mit einem begleitenden Indikatorterm	8.162	-
Übereinstimmungen gültige UniProt / extrahierte UniProt Zugangsnummern mit einem begleitenden Indikatorterm	8.010	58,7% (98,1%)
Anzahl extrahierte UniProt Zugangsnummern mit einem begleitenden Indikatorterm Zeichenabstand < 100	4.743	-
Übereinstimmungen gültige UniProt / extrahierte UniProt Zugangsnummern mit einem begleitenden Indikatorterm Zeichenabstand < 100	4.618	33,8% (97,4%)

Der Vergleich zeigt, dass 94,5% der extrahierten UniProt Zugangsnummern gültigen UniProt Zugangsnummern entspricht. Die Einschränkungen durch die Bedingung eines Indikatorterms begrenzt die Ergebnismenge der extrahierten Zugangsnummern stark (Reduktion um 40,3%), erreicht aber nur eine Verbesserung der relativen Übereinstimmung von 3,6 Prozentpunkte bzw. 2,9 Prozentpunkte, wenn ein Zeichenabstand von 100 Zeichen zwischen Zugangsnummer und Indikatorterm eingehalten werden muss (Reduktion der Ergebnismenge 65,3%). Der Anwendung einer Filterung durch einen Indikatorterm muss eine Abwägung der Reduktion der Ergebnismenge gegen die Reduktion der falsch positiven Ergebnisanteile erfolgen.

Vergleich mit den Einträgen in BRENDA

Ein weiterer Abgleich der extrahierten Zugangsnummern mit gültigen Zugangsnummern, bestand aus dem Vergleich der bereits in der BRENDA Datenbank enthaltenen manuell extrahierten Zugangsnummern (Tabelle 3.23) mit den durch die regelbasierte Suche extrahierten Zugangsnummern. Verglichen wurden alle Zugangsnummern, die auf die selbe Ursprungsreferenz zurückgeführt werden konnten (Tabelle 3.24).

Tabelle 3.23: Die Anzahl der manuell annotierten Zugangsnummern in BRENDA (Stand Januar 2011) verglichen mit den Ergebnissen der regelbasierten Suche, aufgeschlüsselt nach Ursprungsdatenbank. Der Vergleich berücksichtigt nicht die Deckungsgleichheit der manuell extrahierten Ursprungsreferenzen mit den Referenzen des Volltextkorpus.

Datenbank	Zugangsnummern BRENDA	Zugangsnummern regelbasierte Suche
UniProt/SwissProt/TREMBL	15.400	13.677
GenBank	226	43.859
EMBL	23	12.640
ohne Quellangabe	12	-

Die Verteilung der Mengenanteile der Zugangsnummern, bei der Betrachtung aufgeschlüsselt nach Datenbanken, unterscheidet sich erheblich. Der Anteil der UniProt/SwissProt/TREMBL Zugangsnummern an der Gesamtmenge der in BRENDA enthaltenen Zugangsnummern beträgt 1,4%. Der Anteil der UniProt/SwissProt/TREMBL Zugangsnummern beträgt 12,2% der Gesamtmenge der unterschiedlichen Zugangsnummern, die aus der regelbasierten Suche extrahiert werden konnten. Um diese Mengenanteilunterschiede zu analysieren, sind die Ursprungsreferenzen manuell extrahierten Zugangsnummern mit den Referenzen des Volltextkorpus zur Deckung gebracht worden (Tabelle 3.24).

Tabelle 3.24: Die Zugangsnummern in BRENDA Einträgen, die Referenzen entstammen, die zugleich Teilmenge des Volltextkorpus waren und der Anteil, der übereinstimmenden Zugangsnummern aus der regelbasierten Suche für die gleiche Referenz.

Zugangsnummern BRENDA Referenz auch in Volltextkorpus	Anzahl der übereinstimmenden Zugangsnummern
10.360	2,373

Es konnten für 23% der in BRENDA annotierten Zugangsnummern, die aus Referenzen stammten und Teilmenge des *Volltextkorpus* waren, Entsprechungen in den, durch die regelbasierte Suche extrahierten Zugangsnummern gefunden werden. Bei der manuellen Annotation der BRENDA Einträge zu Zugangsnummern sind die Annotatoren angehalten, gefundene Zugangsnummern aus anderen Datenbanken, wenn möglich auf eine UniProt/SwissProt/TREMBL Zugangsnummer zurückzuführen. Deshalb ist die verminderte Übereinstimmung erklärbar, denn der Anteil an der Gesamtmenge der gefundenen Zugangsnummern für GenBank beträgt 40% (Tabelle 3.20), wohingegen der Anteil der Einträge von GenBank an allen manuellen extrahierten Zugangsnummern in BRENDA nur 1,4% (Tabelle 3.23) beträgt.

Ringschluss: Zugangsnummer - Enzymsequenz - Krankheit

Eine Stichprobe von 961 zufällig ausgewählten Zugangsnummern (Tabelle 3.25) aus der Gesamtmenge der 112.522 extrahierten Zugangsnummern wurde daraufhin ausgewertet, ob es möglich ist extrahierte Zugangsnummern eindeutig einer EC Nummer und einer Aminosäure- bzw. Nukleiotidsequenz zuzuordnen.

Tabelle 3.25: Die Zusammensetzung der ausgewerteten Stichprobe der extrahierten Zugangsnummern aufgeschlüsselt nach Ursprungsdatenbank.

Datenbank	Anzahl Zugangsnummern
GenBank	327
PDB	279
NCBI RefSeq	98
UniProt/SwissProt/TREMBL	100
EMBL	102
DDBJ	55
∑ alle Datenbanken	961

Tabelle 3.26: Die Verknüpfungen der extrahierten Zugangsnummern zu Einträgen in den Datenbanken UniProt und BRENDA. Die aufgeführten Zahlen basieren auf einem Abgleich [94] der Stichprobe der extrahierten Zugangsnummer mit den Datenbanken UniProt (Stand April 2011) und BRENDA (Stand Januar 2011) sowie den Ergebnissen der Identifikation der Enzym- und Krankheits-Kookkurrenzen und deren Klassifizierung in die Kategorien causal interaction, ongoing research, diagnostic usage und therapeutic application.

	Art der Beziehungen	Anzahl der Entsprechungen
A	Anzahl der extrahierten Zugangsnummern, für die ein UniProt Zugangsnummer durch einen Abgleich hergeleitet werden konnte.	910
B	Anzahl der extrahierten Zugangsnummern, für in UniProt eine EC Nummer eingetragen ist.	393
C	Anzahl der extrahierten Zugangsnummern, für deren extrahierte Referenz in BRENDA eine EC Nummer zugeordnet ist.	760
D	Anzahl der extrahierten Zugangsnummern die gleichzeitig Bedingung **A**, **B** und **C** erfüllen.	306
E	Anzahl der Referenzen, aus **D**, die gleichzeitig in den Ergebnissen der Identifikation der Kookkurrenz für eine Enzym und Krankheit sind.	38
F	Anzahl der Referenzen aus **E**, die gleichzeitig in den Ergebnissen der Klassifizierung ein oder mehreren Kategorien zugeordnet wurden.	32

An einem Beispiel aus den 32 Referenzen (Tabelle 3.26, F) soll gezeigt werden, dass die Information über eine Zugangsnummer auch eine relevante Zusatzinformation ist, ergänzend zur Klassifizierung der Referenz in die Kategorien *causal interaction*, *ongoing research, diagnostic usage* oder *therapeutic application*.

Die Referenz mit der PubMed ID *19694615* und dem Titel „The crystal structure of caspase-6, a selective effector of axonal degeneration.“ [95] war Teil der ausgewerteten Stichprobe. Diese Referenz wurde bei der Klassifizierung den Kategorien t*herapeutic application* und *causal interaction* zugeordnet. Bei der regelbasierten Suche konnten insgesamt 17 Zugangsnummern aus dieser Referenz extrahiert werden. Darunter auch *P55212,* der UniProt Zugangsnummer für den Eintrag der humanen Caspase-6 und *2WDP*, der PDB Zugangsnummer für die Kristallstruktur der humanen Caspase-6. Diese Beispielreferenz gibt Aufschluss über die Kristallstruktur der humanen Caspase-6 und enthält ausführliche Informationen zu ihren untersuchten kinetischen Parametern, wie k_{cat} und K_m.

Das Enzym Caspase-6 (EC Nummer 3.4.22.59) ist als Effektor Teil der Signalkaskade des Zelltods durch Apoptose [96]. Der Prozess der Apoptose steht in enger Verbindung mit pathophysiologischen Vorgängen, die mit neurodegenerativen Erkrankungen einhergehen [97], wie z.B. Morbus Alzheimer oder Morbus Huntington. Im Inhalt der Beispielreferenz wird darauf verwiesen, dass die genaue Aufklärung der Struktur der Caspase-6 zum Ziel hatte zur Entdeckung von spezifischen Inhibitoren der Caspase-6 beizutragen, weil das Enzym ein potentielles molekulares Wirkziel im Rahmen der Therapie von Morbus Huntington ist, aber auch von antineurodegenerativen Therapeutika im Allgemeinen [95].

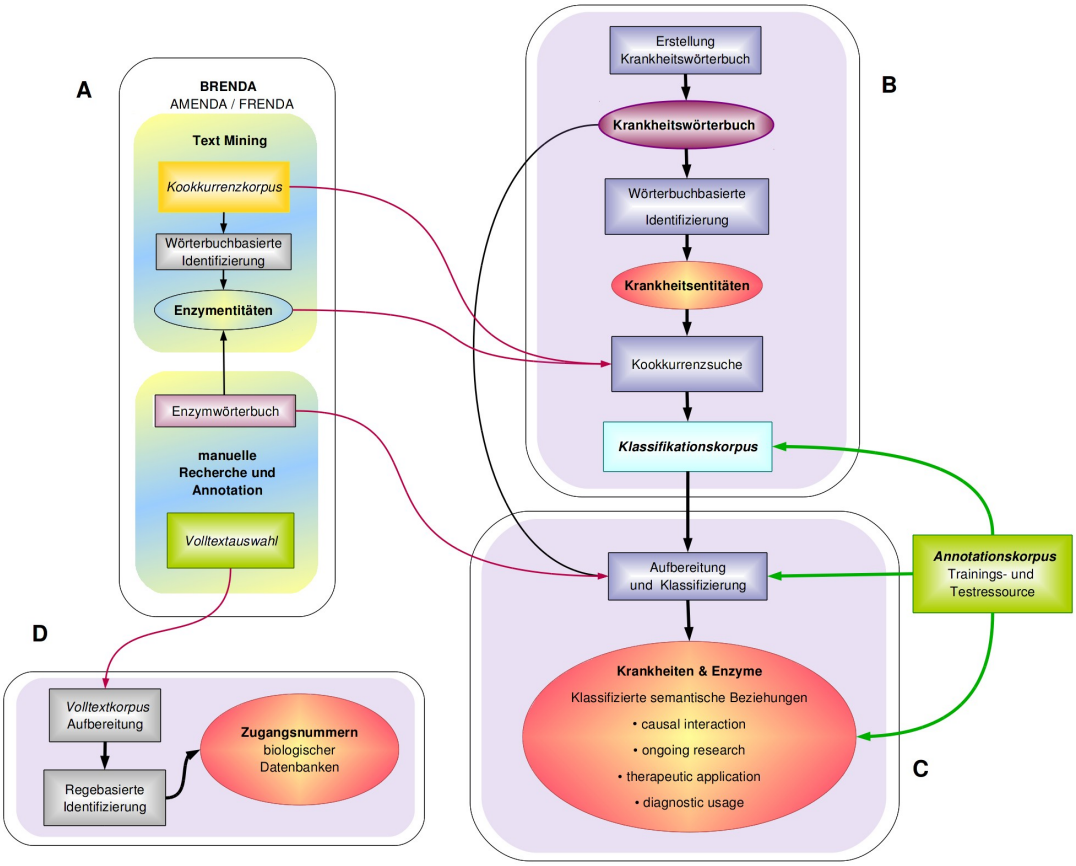
Dieses Beispiel zeigt, dass durch die regelbasierte Suche Zugangsnummern gefunden werden konnten, die für eine organismusspezifische Sequenz einer EC Nummer stehen. Darüber hinaus kann festgehalten werden, dass die zusätzliche Erfassung der Zugangsnummern, begleitend zur Identifikation der Enzym- und Krankheitskookkurrenzen und deren Klassifizierung, den Weg zu relevanten Zusatzinformationen weisen kann.

3.6. Programmabläufe und Laufzeiten

Die im Rahmen dieser Arbeit implementierten Programmabläufe sind in mehrere Elemente unterteilt. Die Implementierung fand unter den in Abschnitt *2.6. Implementation* (ab Seite 44) erläuterten Gesichtspunkten statt. Eine graphische Übersicht über die entwickelten Hauptroutinen befindet sich in Abbildung 3.6. Im Folgenden werden kurz die Abläufe und die benötigte Laufzeit der Programme beschrieben.

Einige Elemente sind nach ihrer Fertigstellung in die regelmäßige Aktualisierung des BRENDA Informationssystems eingebunden worden (Abbildung 3.6, B+C). Es handelt sich dabei um die Identifizierung von Krankheitsentitäten und deren Kookkurrenzen mit Enzymentitäten (Abbildung 3.6, B) sowie die Klassifizierung der semantischen Relationen (Abbildung 3.6, C). An mehreren Stellen (Abbildung 3.6, grüne Pfeile) wird der *Annotationskorpus* benötigt. Der *Annotationskorpus* dient zum einen an diesen Stellen in Teilen zur Überprüfung der Ergebnisse, zum anderen in Teilen zu Training der SVM. Das Krankheitswörterbuch wird sowohl bei der Identifikation von Krankheitsentitäten verwendet als auch bei der Vorverarbeitung der Eingabedaten der Klassifizierung. Am Ende der Verarbeitung dieser beiden Routinen werden zwei Datenbanktabellen erzeugt, die bereits fertig aufbereitet in die übrige BRENDA Datenbank übernommen werden können.

Abbildung 3.6: Eine schematische Darstellung der Programmabläufe. Die Identifizierung von Krankheitselementen (B) geht der Klassifizierung von semantischen Relationen (C) voran. Beide Routinen sind inzwischen in die BRENDA Aktualisierungsroutine (A) integriert worden. Die BRENDA Aktualisierungsroutine (A) wurde hier nur rudimentär in den Teilen skizziert, die für die im Rahmen dieser Arbeit entwickelten Routinen verknüpft sind. Die Identifizierung von Zugangsnummern biologischer Datenbanken (D) ist zur Zeit kein Bestandteil der halbjährlichen Aktualisierung des BRENDA Informationssystems.



Sowohl die Routine der Identifizierung von Krankheitsentitäten und deren Kookkurrenzen mit Enzymentitäten (Abbildung 3.6, B) als auch die Klassifizierung der semantischen Relationen (Abbildung 3.6, C) wurden so implementiert, dass sie auf Ebene der Hardware parallelisiert auf mehreren Knoten eines Computerclusters ausgeführt werden können. Die Laufzeiten der in Abbildung 3.6 skizzierten Routinen B-C ist in Tabelle 3.27 wider gegeben.

Tabelle 3.27: Die Laufzeiten für die jeweiligen Programmroutinen.

Programmroutine	Parallelisiert auf Hardwareebene	Hardware	Laufzeit
Identifizierung von Krankheitsentitäten Kookkurrenzsuche mit Enzymentitäten (Abbildung 3.6, B)	ja	8 Clusterknoten je 12 GB RAM Speicher 2 CPU / 2400 MHz	~ 3 h [93]
Klassifizierung der semantischen Relationen (Abbildung 3.6, C)	ja	8 Clusterknoten je 12 GB RAM Speicher 2 CPU / 2400 MHz	~20 h [93]
Identifizierung von Zugangsnummern biologischer Datenbanken (Abbildung 3.6, D)	nein	1 Clusterknoten je 16 GB RAM Speicher 2 CPU / 1999 MHz	~ 0.6 h

Die auf der Hardwareebene parallelisierten Routinen (Tabelle 3.27, B+C) nutzen acht Knoten eines Rechnerclusters und werden darauf als Einzelthreadroutine gestartet, das heißt die Anwendung nutzt nicht die gesamten Ressourcen der aufgeführten Hardware (Tabelle 3.27, Hardware). Bislang ergeben sich bei der Gesamtlaufzeit der halbjährlichen Aktualisierung des BRENDA Informationssystems keine Laufzeitengpässe. Sofern es ein Zuwachs an zu verarbeitenden Eingabedaten erforderlich machen würde, könnten die Laufzeiten aller Programme (Tabelle 3.27, B-D) durch ihre Anpassung zu Multithreadingroutinen (Parallelisierung auf der Ebene der Software) noch einmal moderat verkürzt werden.

3.7. Fazit und Ausblick

Durch die im Rahmen dieser Arbeit neu- und weiterentwickelten Methoden konnten automatisiert enzymbezogene Informationen aus der wissenschaftlichen Primärliteratur gewonnen werden. Die Identifikation von Referenzen, die Enzym- und Krankheitsentitäten enthalten, erreicht eine gute bis sehr gute Qualität (F_1 Maß 0,89, siehe Tabelle 3.7). Allerdings gelingt es Text Mining gestützten Verarbeitungen bisher nicht den Qualitätsgrad einer manuellen Annotation von Daten gleich zu kommen. Die Suche von Kookkurrenzen in Sätzen bzw. Titeln vernachlässigt zwar satzübergreifende

Kookkurrenzen, jedoch geht die zu erwartende Verbesserung der Vollständigkeit mit einer potentiell erheblichen Verlängerung der Rechenlaufzeit (Mehraufwand der Disambiguierung und Abwägung der Relevanz jeder einzelnen Kookkurrenz) und einer Verschlechterung der Präzision [98] einher. Der vormals in der BRENDA verwendete Ansatz zur konzeptbasierten Identifikation von kookkurrierenden Enzymen und Krankheiten [60,61] arbeitete bereits auf einem qualitativ hohem Niveau, unterlag aber Einschränkungen der Laufzeit im Hinblick auf die Größe des zu analysierenden Textkörpers. Er konnte durch die Anwendung der wörterbuchbasierten Identifikation, die diesen Einschränkungen nicht unterliegt, erfolgreich ersetzt werden. Die Möglichkeit der nachgelagerten Klassifizierung von semantischen Relationen von kookkurrierenden Enzym- und Krankheitsentitäten fehlte bis dahin und ist erst durch die in dieser Arbeit entwickelten Methoden dafür erschlossen worden.

Die kombinierten Ergebnisse, gewonnen aus der Erfassung von Kookkurrenzen und der Klassifizierung der semantischen Relationen, ermöglichen durch ihre Integration in das BRENDA Informationssystem die gezielte Suche nach relevanten Referenzen, die Informationen zu Enzymen und Krankheiten enthalten. Die Verweise zu den entsprechenden Referenzen sind übersichtlich und nachvollziehbar nach der jeweiligen EC Nummer und dem Krankheitsnamen geordnet, abrufbar. Durch die Nutzung der Suche innerhalb der Kategorien *causal interaction*, *ongoing research, diagnostic usage* und *therapeutic application* kann ein Überblick über Referenzen gewonnen werden, die mit hoher Wahrscheinlichkeit wichtige Aussagen zu kausalen Verknüpfungen von Enzymen und Krankheiten, sowie der möglichen diagnostischen Verwendung und therapeutischen Implikation von Enzymen enthalten können. Darüber hinaus kann eine große Auswahl an Referenzen gefunden werden, die einen Spiegel des aktuellen Stands der Forschung an einem Enzym und seiner Verbindung zu einer Krankheit bietet.

Sowohl in der Kookkurrenzsuche als auch in der Klassifizierung der semantischen Beziehungen von Krankheits- und Enzymentitäten ist der Satz respektive der Titel die Basis der Beurteilung. Dies geschieht unter der Annahme, dass die Nennung einer Krankheit und eines Enzyms in einem Titel oder der Kurzzusammenfassung auch eine inhaltliche Relevanz des Artikels widerspiegelt. Die Kurzzusammenfassung enthält für gewöhnlich die Essenz eines Artikels. Bereits vor der Publikation eines Artikels, nämlich bei der Auswahl durch Verlagseditoren, spielt die Kurzzusammenfassung eine wichtige Rolle [99]. Die meisten klinisch-medizinischen Zeitschriften verlangen eine standardisierte Form der Kurzzusammenfassung, die prägnante Informationen zu den Zielen, Methoden und Ergebnissen enthält [100]. Darüber hinaus fordern die Verlage, dass eine Kurzzusammenfassung keine Informationen enthält, die nicht auch dem Gesamtartikel entnommen werden können. Dennoch ist nicht ausgeschlossen, dass es zu

Unrichtigkeiten und Diskrepanzen zwischen der Kurzzusammenfassung und dem Gesamtartikel kommen kann [101]. Diese Art von Falschinformation kann bei einer automatisierten Auswertung von Kurzzusammenfassung mit Text Mining Methoden nicht umgangen werden, wenn sie selbst in der Phase vor dem Erscheinen eines Artikels, bei intensiver Prüfung und Beurteilung durch Editoren und Gutachter nicht aufgefallen ist.

Für die überwiegende Menge der Artikel scheint die Folgerung zulässig, dass das Vorhandensein von Enzym- und Krankheitsentitäten in dem Titel bzw. der Kurzzusammenfassung, auch auf deren Relevanz im Gesamtartikel schließen lässt. Es ist aber davon auszugehen, dass ein großes Problem der Analyse von Kurzzusammenfassungen das nicht Erfassen relevanter Entitäten und Aussagen ist, die nur im Volltextartikel aufgeführt sind. Zum einen könnte dies mit einer Ausweitung der Verarbeitung auf Volltextartikel gelöst werden, zum anderen mit der Erfassung von Informationen wissenschaftlicher Artikel schon im Vorfeld ihrer Veröffentlichung durch Autoren und Editoren. Durch die präpublikatorische Erfassung wäre die Suche nach enthaltenen Entitäten und den Aussagen über deren Verbindung zueinander trivial. Gerade im Hinblick auf die Entwicklung in den Lebenswissenschaften hin zu einer semantischen Vernetzung über das Internet von Inhalten unterschiedlichster Provenienz, die unter dem Begriff „semantisches Web“ [102] zusammengefasst wird, wäre diese Vorgehensweise zunehmend dringlicher [48,103].

Eine Methode dieser Arbeit wurde bereits auf einen Korpus mit dem Inhalt von Volltextartikeln angewendet. Die Identifizierung von Zugangsnummern, die zumeist erst im Volltext erwähnt werden, konnte sehr effizient umgesetzt werden, so dass ein relativ großer Textkorpus in kurzer Zeit ausgewertet werden konnte (Tabelle 3.27). Die überwiegende Mehrzahl (94,5%) aller UniProt zugeordneten Zugangsnummern konnten dabei als gültige UniProt Zugangsnummern bestätigt werden. Durch die automatische Auswertung von Referenzen und die Verknüpfung aller dadurch erhaltenen Informationen ist es gelungen einen Ringschluss von einer Krankheit, die im Zusammenhang zur Aktivität eines Enzyms steht, auf den Verweis einer organismusspezifischen Sequenz dieses Enzyms sowie seiner organismusspezifischen Struktur und wiederum auf die Rolle des Enzyms als potentielles Wirkziel bei der Therapie dieser Krankheit zu vollziehen.

Anhang

A. MeSH Kategorien Englisch

Tabelle Anhang 1: Eine Auflistung aller Subkategorien und deren Namen der Kategorie C (Diseases) der MeSH Hierarchie mit der englischen Originalbezeichnung.

Name der Kategorie	Kategorie
Bacterial Infections and Mycoses	[C01]
Virus Diseases	[C02]
Parasitic Diseases	[C03]
Neoplasms	[C04]
Musculoskeletal Diseases	[C05]
Digestive System Diseases	[C06]
Stomatognathic Diseases	[C07]
Respiratory Tract Diseases	[C08]
Otorhinolaryngologic Diseases	[C09]
Nervous System Diseases	[C10]
Eye Diseases	[C11]
Male Urogenital Diseases	[C12]
Female Urogenital Diseases and Pregnancy Complications	[C13]
Cardiovascular Diseases	[C14]
Hemic and Lymphatic Diseases	[C15]
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	[C16]
Skin and Connective Tissue Diseases	[C17]
Nutritional and Metabolic Diseases	[C18]
Endocrine System Diseases	[C19]
Immune System Diseases	[C20]
Disorders of Environmental Origin	[C21]
Animal Diseases	[C22]
Pathological Conditions, Signs and Symptoms	[C23]
Occupational Diseases	[C24]
Substance-Related Disorders	[C25]
Wounds and Injuries	[C26]

B. Liste der verwendeten statischen Stoppwörter

Tabelle Anhang 2: Stoppwörter und die Frequenz ihres Auftretens in den PubMed Kurzzusammenfassungen.

Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz
a	+	+	33.342.596	be	+	+	6.891.199	definitely	+	-	9.492	gets	+	+	4.531
a's	+	+	618	became	+	+	130.112	described	+	+	603.465	getting	+	+	13.728
able	+	-	258.835	because	+	-	737.887	did	+	-	1.248.880	given	+	-	551.978
about	+	+	924.332	become	+	+	193.520	didn't	+	-	4.607	gives	+	-	60.344
above	+	+	243.481	becomes	+	+	60.736	different	+	+	1.917.976	go	+	+	27.764
according	+	+	372.240	becoming	+	+	41.493	do	+	-	400.104	goes	+	+	8.693
accordingly	+	-	34.796	been	+	-	3.116.471	doesn't	+	-	3.186	going	+	+	13.658
across	+	+	242.899	before	+	+	825.745	doing	+	-	17.866	gone	+	+	3.931
actually	+	+	36.006	beforehand	+	+	1.556	don't	+	-	8.912	got	+	+	10.295
after	+	+	4.505.338	behind	+	+	29.266	downwards	+	+	1.181	gotten	+	+	387
afterwards	+	+	9.443	being	+	-	520.826	during	+	+	3.174.219	had	+	-	2.849.717
against	+	-	803.949	believe	+	+	48.031	each	+	+	1.166.842	hadn't	+	-	199
ain't	+	-	182	below	+	+	161.330	edu	+	+	372	happens	+	-	4.743
all	+	+	2.655.156	beside	+	+	6.038	eg	+	+	5.906	hardly	+	-	13.023
allow	+	-	151.877	besides	+	+	44.758	eight	+	+	297.199	has	+	-	2.945.562
allows	+	+	167.756	best	+	-	227.456	either	+	+	907.615	hasn't	+	-	365
almost	+	+	218.210	better	+	+	408.015	else	+	+	3.361	have	+	-	3.958.430
alone	+	+	391.240	between	+	+	3.801.734	enough	+	+	54.122	haven't	+	-	477
along	+	+	225.977	beyond	+	+	83.720	entirely	+	-	34.500	having	+	-	270.229
also	+	+	2.734.997	both	+	+	3.040.017	especially	+	+	311.992	he	+	+	126.639
although	+	+	939.449	brief	+	+	87.466	et	+	+	155.968	he's	+	-	218
always	+	+	113.000	but	+	+	3.320.744	etc	+	+	14.685	hello	+	+	91
am	+	-	36.411	by	+	+	14.456.299	even	+	+	445.088	help	+	-	191.799
among	+	+	1.201.656	c'mon	+	-	4	ever	+	+	25.339	hence	+	-	78.029
amongst	+	+	25.442	c's	+	+	282	every	+	+	176.979	her	+	+	96.529
an	+	+	7.345.908	came	+	+	22.259	everyone	+	+	3.380	here	+	+	485.360
and	+	+	67.503.626	can	+	-	2.291.937	everything	+	+	2.416	here's	+	-	855
another	+	+	243.600	cannot	+	-	142.791	everywhere	+	+	1.510	hereafter	+	+	841
any	+	+	690.171	cant	+	-	234	ex	+	+	43.450	hereby	+	+	2.060
anybody	+	+	237	causes	+	-	298.971	exactly	+	+	14.869	herein	+	+	39.601
anyhow	+	+	203	certain	+	-	253.965	example	+	+	123.487	hereupon	+	+	13
anyone	+	+	2.067	certainly	+	-	11.077	except	+	+	168.879	hers	+	+	429
anything	+	+	3.058	changes	+	-	1.614.418	far	+	+	135.511	herself	+	+	1.724
anyway	+	+	707	clearly	+	-	154.533	few	+	+	294.819	hi	+	+	11.639
anyways	+	+	3	co	+	+	52.653	fifth	+	+	34.038	him	+	+	8.824
anywhere	+	+	2.607	com	+	+	2.632	first	+	+	1.341.844	himself	+	+	4.353
apart	+	+	42.907	come	+	+	31.739	five	+	+	583.807	his	+	+	162.741
appreciate	+	-	4.528	comes	+	-	14.911	followed	+	+	536.758	hither	+	+	37
are	+	-	7.090.637	concerning	+	-	138.236	following	+	+	1.075.302	hopefully	+	+	4.412
aren't	+	-	909	consequently	+	+	65.208	follows	+	+	80.272	how	+	+	379.819
around	+	+	178.915	consider	+	-	78.927	for	+	+	17.905.522	howbeit	+	+	1
as	+	+	8.920.305	contain	+	+	166.369	former	+	+	73.271	however	+	+	1.691.549
aside	+	+	4.542	containing	+	-	505.855	formerly	+	+	7.936	i	+	+	705.058
ask	+	+	9.124	contains	+	-	167.807	forth	+	+	5.960	i'd	+	+	109
asking	+	+	6.643	corresponding	+	-	283.583	four	+	+	874.780	i'll	+	+	130
associated	+	-	1.891.034	could	+	-	1.166.522	from	+	+	9.258.265	i'm	+	+	927
at	+	+	7.368.192	couldn't	+	-	726	further	+	-	856.918	i've	+	-	607
away	+	+	30.984	course	+	+	385.546	furthermore	+	+	361.496	ie	+	+	14.096
awfully	+	-	11	currently	+	+	168.361	get	+	+	28.208	if	+	+	622.253

Tabelle Anhang 3: Fortsetzung von Tabelle Anhang 2

Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz
ignored	+	-	9.151	might	+	+	408.825	ours	+	+	2.049	serious	+	+	129.749
immediate	+	+	122.462	moreover	+	+	241.155	ourselves	+	+	2.130	seriously	+	+	13.816
in	+	+	60.536.857	mostly	+	+	80.987	out	+	+	636.093	seven	+	+	307.658
inasmuch	+	+	3.315	my	+	+	19.338	outside	+	+	71.750	several	+	+	814.526
inc	+	+	23.300	name	+	+	21.862	over	+	+	974.210	shall	+	-	4.676
indeed	+	-	50.333	namely	+	+	63.745	overall	+	+	469.191	she	+	+	84.704
indicate	+	-	654.161	nd	+	+	7.291	own	+	+	102.206	should	+	-	854.708
indicated	+	-	460.017	near	+	+	158.795	particularly	+	-	283.681	shouldn't	+	-	418
indicates	+	-	191.851	nearly	+	+	118.420	per	+	+	800.206	since	+	-	464.487
inner	+	+	103.676	necessary	+	-	291.467	perhaps	+	+	50.029	six	+	+	494.675
insofar	+	+	1.716	need	+	-	374.978	placed	+	+	131.942	so	+	+	295.603
instead	+	+	80.570	needs	+	-	157.100	please	+	+	2.030	some	+	+	1.216.042
into	+	+	1.774.509	neither	+	-	170.862	plus	+	+	208.392	somebody	+	+	243
inward	+	+	24.956	never	+	-	68.386	possible	+	+	765.653	somehow	+	+	2.318
is	+	-	13.438.037	nevertheless	+	+	66.576	presumably	+	+	53.543	someone	+	+	4.205
isn't	+	-	1.800	new	+	+	1.480.838	probably	+	-	235.284	something	+	+	6.642
it	+	+	3.031.125	next	+	+	79.950	provides	+	-	296.225	sometime	+	+	1.121
it'll	+	+	18	nine	+	+	216.023	que	+	+	1.174	sometimes	+	+	59.977
it's	+	-	11.847	no	+	-	2.722.633	quite	+	+	65.469	somewhat	+	+	40.030
its	+	-	1.991.957	nobody	+	+	640	qv	+	+	104	somewhere	+	+	1.210
itself	+	+	103.216	non	+	-	44.337	rather	+	+	264.656	soon	+	+	33.280
just	+	+	79.705	none	+	-	170.463	rd	+	+	9.381	sorry	+	+	191
keep	+	+	20.952	noone	+	-	9	re	+	+	19.595	specified	+	-	25.390
kept	+	+	43.674	nor	+	-	195.810	really	+	+	13.559	specify	+	-	9.441
know	+	+	39.783	normally	+	+	91.681	reasonably	+	-	13.576	specifying	+	-	5.525
known	+	-	618.764	not	+	-	5.466.439	regarding	+	-	216.294	still	+	+	311.606
knows	+	+	1.317	nothing	+	+	7.229	regardless	+	+	77.794	sub	+	+	2.545
last	+	+	209.300	novel	+	+	520.471	regards	+	+	16.920	such	+	+	1.405.742
lately	+	+	1.997	now	+	+	231.051	relatively	+	+	295.633	sup	+	+	932
latter	+	+	175.541	nowhere	+	+	602	respectively	+	-	1.050.918	sure	+	-	4.305
latterly	+	+	137	obviously	+	+	19.993	right	+	+	413.198	t's	+	+	184
least	+	+	478.085	of	+	+	99.377.218	said	+	+	15.805	take	+	+	95.857
less	+	+	1.298.051	off	+	+	51.516	same	+	+	842.412	taken	+	+	297.774
lest	+	+	444	often	+	+	403.216	saw	+	+	10.489	tell	+	+	5.826
let	+	-	8.245	oh	+	+	12.733	say	+	+	10.187	tends	+	-	15.587
let's	+	-	1.639	ok	+	+	1.993	saying	+	+	1.121	th	+	+	27.128
like	+	+	157.037	okay	+	+	120	says	+	+	4.314	than	+	+	4.160.292
liked	+	+	1.464	old	+	+	229.499	second	+	+	556.237	thank	+	+	1.114
likely	+	+	415.317	on	+	+	10.625.940	secondly	+	+	13.281	thanks	+	+	6.381
little	+	+	292.511	once	+	+	121.027	see	+	+	56.043	thanx	+	+	1
look	+	+	29.911	one	+	+	2.208.951	seeing	+	+	6.120	that	+	+	13.420.503
looking	+	+	16.715	ones	+	+	74.503	seem	+	+	105.455	that's	+	+	1.212
looks	+	-	7.811	only	+	+	1.919.113	seemed	+	+	51.073	thats	+	-	10
ltd	+	+	11.583	onto	+	+	70.848	seeming	+	+	757	the	+	+	111.748.671
mainly	+	-	246.638	or	+	+	7.723.943	seems	+	-	189.431	their	+	+	2.626.880
may	+	-	2.958.142	other	+	+	2.289.274	seen	+	+	498.303	theirs	+	-	536
maybe	+	+	3.344	others	+	+	147.201	self	+	+	37.764	them	+	+	377.391
me	+	+	16.502	otherwise	+	+	53.378	selves	+	+	579	themselves	+	+	59.420
meanwhile	+	+	10.484	ought	+	+	3.469	sensible	+	-	2.856	then	+	+	486.782
merely	+	+	12.484	our	+	+	1.385.160	sent	+	+	27.654	thence	+	-	495

Tabelle Anhang 4: Fortsetzung Tabelle Anhang 3

Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz	Stoppwort	A	B	Frequenz
there	+	+	1.889.110	used	+	-	2.357.267	willing	+	+	8.835
there's	+	-	1.181	useful	+	-	433.793	wish	+	+	9.735
thereafter	+	+	52.010	uses	+	-	73.268	with	+	+	24.081.891
thereby	+	+	94.230	using	+	+	2.907.588	within	+	+	1.208.078
therefore	+	+	529.839	usually	+	+	203.069	without	+	+	1.098.991
therein	+	+	2.132	value	+	-	568.536	won't	+	-	1.213
theres	+	-	9	very	+	+	596.881	wonder	+	+	1.018
thereupon	+	+	246	via	+	+	427.051	would	+	-	400.525
these	+	+	5.310.043	viz	+	+	3.385	wouldn't	+	-	198
they	+	+	1.071.847	vs	+	+	511.373	yes	+	+	2.842
they'd	+	-	30	want	+	+	12.408	yet	+	+	192.578
they'll	+	+	88	wants	+	+	2.336	you	+	+	34.773
they're	+	+	756	was	+	-	18.071.513	you'd	+	-	52
they've	+	+	204	wasn't	+	-	868	you'll	+	-	281
think	+	+	15.697	way	+	+	179.052	you're	+	+	1.064
third	+	+	245.403	we	+	+	5.244.193	you've	+	-	352
this	+	+	7.825.677	we'd	+	-	42	your	+	+	36.818
thorough	+	+	24.159	we'll	+	-	269	yours	+	+	172
thoroughly	+	+	11.120	we're	+	-	524	yourself	+	+	1.343
those	+	+	1.501.509	we've	+	-	459	yourselves	+	+	29
though	+	+	121.378	welcome	+	+	2.577	zero	+	+	41.409
three	+	+	1.631.365	well	+	+	1.357.724				
through	+	+	995.378	went	+	+	11.764				
thru	+	+	188	were	+	-	15.674.924				
thus	+	+	739.468	weren't	+	-	316				
tis	+	+	1.500	what	+	+	202.825				
to	+	+	35.447.912	what's	+	-	7.400				
together	+	+	293.867	whatever	+	+	10.101				
too	+	+	70.760	when	+	+	1.936.603				
took	+	-	54.857	whence	+	+	271				
toward	+	+	186.044	whenever	+	+	14.765				
towards	+	-	150.869	where	+	+	419.456				
tried	+	+	19.502	where's	+	+	308				
tries	+	+	2.743	whereafter	+	+	526				
truly	+	-	10.255	whereas	+	+	770.396				
try	+	+	12.555	whereby	+	+	25.151				
trying	+	+	8.653	wherein	+	+	7.599				
twas	+	+	74	whereupon	+	+	808				
twice	+	+	111.106	wherever	+	+	2.148				
two	+	+	3.105.458	whether	+	+	700.402				
un	+	+	3.005	which	+	+	4.015.295				
under	+	+	946.176	while	+	+	1.017.992				
unfortunately	+	+	20.220	who	+	+	1.383.147				
unless	+	+	24.010	who's	+	+	2.324				
until	+	+	192.394	whoever	+	+	79				
unto	+	+	212	whole	+	+	252.750				
up	+	+	549.623	whom	+	+	157.545				
upon	+	+	364.486	whose	+	+	183.192				
us	+	+	209.654	why	+	+	65.619				
use	+	-	1.748.715	will	+	+	597.861				

C. Auflistung der abgeleiteten Formatierungsregeln

Die Identifizierung und Zuordnung der Zugangsnummern von wissenschaftlichen Datenbanken basiert auf der Analyse der alphanumerischen Struktur. Die Formatierungsvorgaben der Datenbanknummern wurden aus den Dokumentationen und Hilfeseiten der jeweiligen Datenbanken entnommen. Sofern die alphanumerische Struktur bekannt ist, kann aufbauend darauf ein regulärer Ausdruck erstellt werden.

- In Tabelle Anhang 5 ist aufgeschlüsselt wie eine Zeichenkette aufgebaut sein darf wenn sie eine Zugangsnummer einer Datenbank ist. Die Namen der Datenbanken sind Tabelle Anhang 6 zu entnehmen und dort unter der entsprechenden Nummer aufgeführt.
- In Tabelle Anhang 6 sind die regulären Ausdrücke aufgelistet, die eine Zeichenkette beschreiben. Die regulären Ausdrücke entsprechen der Struktur der Zeichenketten, die in Tabelle Anhang 5 definierten sind.

Tabelle Anhang 5: Auflistung der zulässigen Zeichen der Zeichenkette einer Zugangsnummer.

Nr.	Stelle der Zeichenkette / erlaubte alphanumerische Zeichen an entsprechender Stelle der Zeichenkette													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0-9	a-z; A-Z; 0-9	a-z; A-Z; 0-9	a-z; A-Z; 0-9											
2	A-N; R-Z	0-9	A-Z	A-Z; 0-9	A-Z; 0-9	0-9									
3	O; P; Q	0-9	A-Z; 0-9	A-Z; 0-9	A-Z; 0-9	0-9									
4	B; F-N; R-U; W	0-9	0-9	0-9	0-9	0-9									
5	B; E	A-Z	A-Z	A-Z	0-9	0-9									
6	C; D; E	0-9	0-9	0-9	0-9	0-9									
7	A; F; V; X; Y; Z	0-9	0-9	0-9	0-9	0-9									
8	C	A-Z	A-Z	A-Z	0-9	0-9									
9	PS		0-9	0-9	0-9	0-9	0-9								
10	CH; CM; DS; EM; EN; EP; EQ; FA; GG; GL; JH	A-Z	0-9	0-9	0-9	0-9	0-9	0-9							
11	A; D; E; H	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9							
12	AA; AC; AD; AE; AF; AH; AI; AQ; AR; AS; AW; AY; AZ; BC; BE; BF; BG; BH; BI; BK; BL; BM; BQ; BT; BU; BV; BZ; CA; CB; CC; CD; CE; CF; CG; CK; CL; CN; CO; CP; CV; CW; CX; CY; CZ; DN; DP; DQ; DR; DT; DU; DV; DW; DX; DY; DZ; EA; EB; EC; ED; EE; EF; EG; EH; EI; EJ; EK; EL; ER; ES; ET; EU; EV; EW; EX; EY; EZ; FC; FD; FE; FF; FG; FH; FI; FJ; FK; FL; GC; GD; GE; GF; GH; GI; GK; GO; GP; GQ; GR; GS; GT; GU; GV; GW; GX; GY; GZ; HJ; HK; HL; HM; HN; HO; HP; HQ; HR; HS; JF; JG; JI	0-9	0-9	0-9	0-9	0-9	0-9	0-9						
13	AB; AG; AK; AP; AT; AU; AV; BA; BB; BD; BJ; BP; BR; BS; BW; BY; CI; CJ; DA; DB; DC; DD; DE; DF; DG; DH; DI; DJ; DK; DL; DM; FS; FT; FU; FV; FW; FX; FY; FZ; GA; HT; HU	0-9	0-9	0-9	0-9	0-9	0-9	0-9							
14	B; F; G; I	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9							
15	AJ; AL; AM; AN; AX; BN; BX; CQ; CR; CS; CT; CU; FB; FM; FN; FP; FQ; FR; GM; GN; HA; HB; HC; HD; HE; HF; HG; HH; HI; JA; JB; JC; JD; JE	0-9	0-9	0-9	0-9	0-9	0-9	0-9							
16	C	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9							
17	AC_; NC_; NG_; NT_; NW_; NS_			0-9	0-9	0-9	0-9	0-9	0-9						
18	NM_|NR_|XM_|XR_			0-9	0-9	0-9	0-9	0-9	0-9						
19	AP_; NP_; XP_; YP_			0-9	0-9	0-9	0-9	0-9	0-9						
20	A; D	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9					
21	B; E	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9					
22	C	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9					
23	ZP_			0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9				
24	A; D	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9				
25	B; E	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9				
26	C	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9				
27	NP_; XP_; YP_			0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
28	NM_; XM_			0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
29	NW_			0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
30	A; D	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
31	B; E	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
32	C	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
33	A	A-Z	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9			
34	NZ_			A-Z	A-Z	A-Z	A-Z	0-9	0-9	0-9	0-9	0-9	0-9	0-9	0-9

Tabelle Anhang 6: Auflistung der Datenbanken, der Art des referenzierten Eintrags und des definierten regulären Ausdrucks zur Suche nach dieser Zugangsnummer in einem Text.

Nr.	Datenbank	Art des Eintrags	Regulärer Ausdruck
33	DDBJ	Sequenz aus der Genomannotation	[A][A-Z][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
6	DDBJ	Nucleotidsequenz	[C-E][0-9][0-9][0-9][0-9][0-9]
13	DDBJ	Nucleotidsequenz	(AB|AG|AK|AP|AT|AU|AV|BA|BB|BD|BJ|BP|BR|BS|BW|BY|CI|CJ|DA|DB|DC|DD|DE|DF|DG|DH|DI|DJ|DK|DL|DM|FS|FT|FU|FV|FW|FX|FY|FZ|GA|HT|HU)[0-9][0-9][0-9][0-9][0-9][0-9]
14	DDBJ	Protein	[ADEH][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9]
5	DDBJ	Whole Genome Shotgun	[EB][A-Z][A-Z][A-Z][0-9][0-9]
21	DDBJ	Whole Genome Shotgun	[BE][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9]
25	DDBJ	Whole Genome Shotgun	[BE][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
31	DDBJ	Whole Genome Shotgun	[BE][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
7	EMBL	Nucleotidsequenz	[AFVX-Z][0-9][0-9][0-9][0-9][0-9]
15	EMBL	Nucleotidsequenz	(AJ|AL|AM|AN|AX|BN|BX|CQ|CR|CS|CT|CU|FB|FM|FN|FP|FQ|FR|GM|GN|HA|HB|HC|HD|HE|HF|HG|HH|HI|JA|JB|JC|JD|JE)[0-9][0-9][0-9][0-9][0-9][0-9]
16	EMBL	Protein	[C][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9]
8	EMBL	Whole Genome Shotgun	[C][A-Z][A-Z][A-Z][0-9][0-9]
22	EMBL	Whole Genome Shotgun	[C][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9]
26	EMBL	Whole Genome Shotgun	[C][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
32	EMBL	Whole Genome Shotgun	[C][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
4	GenBank	Nucleotidsequenz	[BF-NR-UW][0-9][0-9][0-9][0-9][0-9]
12	GenBank	Nucleotidsequenz	(AA|AC|AD|AE|AF|AH|AI|AQ|AR|AS|AW|AY|AZ|BC|BE|BF|BG|BH|BI|BK|BL|BM|BQ|BT|BU|BV|BZ|CA|CB|CC|CD|CE|CF|CG|CK|CL|CN|CO|CP|CV|CW|CX|CY|CZ|DN|DP|DQ|DR|DT|DU|DV|DW|DX|DY|DZ|EA|EB|EC|ED|EE|EF|EG|EH|EI|EJ|EK|EL|ER|ES|ET|EU|EV|EW|EX|EY|EZ|FC|FD|FE|FF|FG|FH|FI|FJ|FK|FL|GC|GD|GE|GF|GH|GJ|GK|GO|GP|GQ|GR|GR|GS|GT|GU|GV|GW|GX|GY|GZ|HJ|HK|HL|HM|HN|HO|HP|HQ|HR|HS|JF|JG|JI)[0-9][0-9][0-9][0-9][0-9][0-9]
11	GenBank	Protein	[ADEH][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9]
20	GenBank	Whole Genome Shotgun	[AD][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9]
24	GenBank	Whole Genome Shotgun	[AD][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
30	GenBank	Whole Genome Shotgun	[AD][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
10	NCBI	Nucleotidsequenz	(CH|CM|DS|EM|EN|EP|EQ|FA|GG|GL|JH)[0-9][0-9][0-9][0-9][0-9][0-9]
17	NCBI RefSeq	Genom Molekül	(AC_|NC_|NG_|NT_|NW_|NS_)[0-9][0-9][0-9][0-9][0-9][0-9]
29	NCBI RefSeq	Genom Molekül	(NW_)[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
18	NCBI RefSeq	mRNA	(NM_|NR_|XM_|XR_)[0-9][0-9][0-9][0-9][0-9][0-9]
28	NCBI RefSeq	mRNA	(NM_|XM_)[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
19	NCBI RefSeq	Protein	(AP_|NP_|XP_|YP_)[0-9][0-9][0-9][0-9][0-9][0-9]
23	NCBI RefSeq	Protein	(ZP_)[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
27	NCBI RefSeq	Protein	(NP_|XP_|YP_)[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
34	NCBI RefSeq	Whole Genome Shotgun	(NZ_)[A-Z][A-Z][A-Z][A-Z][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]
1	PDB	Protein- und Nucleinsäurestrukturen	[0-9][a-zA-Z0-9][a-zA-Z0-9][a-zA-Z0-9]
9	PROSITE	Proteindomänen (functional site)	PS[0-9][0-9][0-9][0-9][0-9]
2	SwissProt	Protein	[A-NR-Z][0-9][A-Z][A-Z0-9][A-Z0-9][0-9]
3	SwissProt	Protein	[OPQ][0-9][A-Z0-9][A-Z0-9][A-Z0-9][0-9]

D. Verwendete Programme

Eine Auflistung des Programms, der Versionsnummer und der jeweiligen Quelle

- bash 3.2+
  http://www.gnu.org/software/bash/bash.html
- Grid Engine 6.2u5
  http://gridengine.sunsource.net
- MySQL 5.1.49
  http://www.mysql.com
- MySQLdb 1.2.2-10+b1
  http://sourceforge.net/projects/mysql-python
- pdftotext 0.12.4
  http://poppler.freedesktop.org
- Python 2.6
  http://python.org
- R 2.11.1
  http://www.r-project.org
- ROCR 1.0-4-1
  http://rocr.bioinf.mpi-sb.mpg.de
- SVMlight 6.02
  http://svmlight.joachims.org
- eric4 4.7.7
  http://eric-ide.python-projects.org
- Debian 6.0 (squeeze)
  http://www.debian.org

Abkürzungsverzeichnis

AE	Amerikanisches Englisch
ATP	Adenosintriphosphat,
AUC	Area under the curve
BE	Britisches Englisch
BRENDA	BRaunschweig ENzyme DAtabase
CPU	Central Processing Unit
DDBJ	DNA DataBank of Japan
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory
GO	Gene Ontology
INSDC	International Nucleotide Sequence Database Collaboration
IUBMB	International Union of Biochemistry and Molecular Biology
MCC	Korrelationskoeffizient nach Matthews
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NLM	United States National Library of Medicine
PDB	Protein Data Bank
PDF	Portable Document Format
RAM	Random Access Memory
ROC	Receiver Operator Characteristics

SVM	Support Vector Machine
tf-idf	Term frequency – inverted document fruquency

Glossar

Akronym

Eine Abkürzung, die sich aus den Anfangsbuchstaben mehrerer Wörter zusammensetzt.

Einzelthreadroutine

Eine nicht parallelisierte Programmroutine. Diese kann aber auf der Hardwareebene parallelisiert ablaufen, sofern sie auf mehreren Rechnern parallel aufgerufen wird.

Entität

Ein konkretes oder abstraktes Objekt, welchem Informationen als Eigenschaften zugeordnet werden können.

Homonym

Ein Ausdruck, der mehrere Begriffe zugleich beschreibt. In dieser Arbeit auch ein Name, der mehr als eine Entität bezeichnet.

Kookkurrenz

In einer übergeordneten Informationseinheit treten zwei lexikalische Einheiten (z.B. identifizierte Entitäten) gemeinsames auf.

Multithreadingroutine

Eine Programmroutine, die auf der Softwareebene parallelisiert abläuft (in Abgrenzung zu Parallelisierung auf Hardwareebene).

N-Gramm

Eine Folge von n Buchstaben oder Worten.

Name

Eine Bezeichnung einer Entität, die auch aus mehreren Worten bestehen kann.

Named Entity Recognition

Die Identifizierung eines Objektes anhand der Nennung seines Namens oder einer Umschreibung für die Entität.

Ontologie

Eine Sammlung von Begriffen, die zudem entsprechend ihrer semantischen und hierachischen Relationen gegliedert sind.

Regulärer Ausdruck

Eine definierte syntaktische Regel, die eine Zeichenkette oder eine Menge von Zeichenketten definiert und so als Filterkriterium oder Schablone dienen kann.

Reintext

Ein unverschlüsselter Text, der ohne weitere Umwandlung für das menschliche Auge lesbar vorliegt.

Stoppwort

Ein Wort mit vermeintlich niedrigem Informationsgehalt.

Support Vector Machine

Mathematisches Verfahren des maschinellen Lernens zur Klassifizierung von Objekten mit unbekannter Klassenzugehörigkeit unter Verwendung von Objekten mit bereits bekannter Klassenzugehörigkeit.

Synonym

Ein Ausdruck, der die gleiche Bedeutung hat wie ein anderes Wort im gleichen Kontext. In dieser Arbeit auch ein Name, der die gleiche Entität beschreibt wie ein anderer Name zur Beschreibung der gleichen Entität.

Text Mining

Die Disziplin der automatischen Informationsgewinnung aus Daten, und hier insbesondere Texten

Textkorpus

Eine Sammlung von Texten deren Auswahl und Zusammenstellung oft zweckorientiert ist.

Abbildungsverzeichnis

Tabellenverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

1. Friboulet, A. und Thomas, D. **Systems Biology-an interdisciplinary approach**. *Biosensors & Bioelectronics 20*, 12 (2005), 2404-2407.

2. Harmston, N., Filsell, W., und Stumpf, M.P.H. **What the papers say: text mining for genomics and systems biology**. *Human Genomics 5*, 1 (2010), 17-29.

3. Cohen, A.M. und Hersh, W. **A survey of current work in biomedical text mining.** *Brief Bioinformatics 6*, 1 (2005), 57-71.

4. Mehler, A. und Wolff, C. **Einleitung: Perspektiven und Positionen des Text Mining [Einführung in das Themenheft Text Mining des LDV-Forum]**. *LDV-Forum 20*, 1 (2005), 1–18.

5. Manning, C.D., Raghavan, P., und Schütze, H. ***Introduction to Information Retrieval***. Cambridge University Press, 2008.

6. Ananiadou, S. und Mcnaught, J. ***Text Mining for Biology And Biomedicine***. Artech House Publishers, 2005.

7. Hand, D.J., Smyth, P., und Mannila, H. ***Principles of data mining***. MIT Press, Cambridge, MA, USA, 2001.

8. Rzhetsky, A., Seringhaus, M., und Gerstein, M. **Seeking a New Biology through Text Mining**. *Cell 134*, 1 (2008), 9-13.

9. Pertsch, E. ***Langenscheidts Handwörterbuch lateinisch-deutsch : auf der Grundlage des Menge-Güthling***. Langenscheidt, Berlin, 1980.

10. **Entität**. *Wiktionary Das freie Wörterbuch*. http://de.wiktionary.org/wiki/Entität.

11. Krallinger, M., Erhardt, R.A.A., und Valencia, A. **Text-mining approaches in molecular biology and biomedicine**. *Drug discovery today 10*, 6 (2005), 439–445.

12. Erhardt, R., Schneider, R., und Blaschke, C. **Status of text-mining techniques applied to biomedical text**. *Drug Discovery Today 11*, 7-8 (2006), 315-325.

13. Jensen, L.J., Saric, J., und Bork, P. **Literature mining for the biologist: from information retrieval to biological discovery**. *Nature Reviews Genetics 7*, 2 (2006), 119–129.

14. Hirschman, L., Morgan, A.A., und Yeh, A.S. **Rutabaga by any other name: extracting biological names**. *Journal of Biomedical Informatics 35*, 4 (2002), 247–259.

15. Webb, E. und International Union of Biochemistry and Molecular Biology. ***Enzyme nomenclature 1992 : Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenc lature and classification of enzymes***. Academic Press, New York, 1992.

16. Ananiadou, S., Kell, D.B., und Tsujii, J. **Text mining and its potential applications in systems biology**. *Trends in Biotechnology 24*, 12 (2006), 571-579.

17. Bussmann, H. ***Lexikon der Sprachwissenschaft : mit 14 Tabellen***. Kröner, Stuttgart, 2008.

18. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., und Garner, H.R. **Knowledge discovery by automated identification and ranking of implicit relationships**. *Bioinformatics (Oxford, England) 20*, 3 (2004), 389-398.

19. Vapnik, V.N. ***The nature of statistical learning theory***. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

20. Voet, D. und Voet, J.G. ***Biochemistry.*** J. Wiley, New York, 1995.

21. Gersting, S.W., Kemter, K.F., Staudigl, M., Messing, D.D., Danecka, M.K., Lagler, F.B., Sommerhoff, C.P., Roscher, A.A., und Muntau, A.C. **Loss of function in phenylketonuria is caused by impaired molecular motions and conformational instability**. *American Journal of Human Genetics 83*, 1 (2008), 5-17.

22. Frei, K., Truong, D.D., und Dressler, D. **Botulinum toxin therapy of hemifacial spasm: comparing different therapeutic preparations**. *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies 13 Suppl 1*, (2006), 30-35.

23. Pschyrembel, W. ***Pschyrembel Klinisches Wörterbuch : [mit 280 Tabellen]***. De Gruyter, Berlin ;;New York, 2002.

24. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., u. a. **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Research*, (2010).

25. Jensen, L.J., Saric, J., und Bork, P. **Literature mining for the biologist: from information retrieval to biological discovery**. *Nature Reviews. Genetics 7*, 2 (2006), 119-129.

26. SEWELL, W. **MEDICAL SUBJECT HEADINGS IN MEDLARS**. *Bulletin of the Medical Library Association 52*, (1964), 164-170.

27. Lindberg, D.A.B. **The National Library of Medicine**. *World Neurosurgery 74*, 1 (2010), 46-48.

28. Stuart, J.N., Powell, T., und Humphreys, B.L. ***The Unified Medical Language System® (UMLS®) Project***. http://www.nlm.nih.gov/mesh/umlsforelis.html, 2006.

29. Powell, T., Srinivasan, S., Nelson, S.J., Hole, W.T., Roth, L., und Olenichev, V. **Tracking meaning over time in the UMLS Metathesaurus**. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, (2002), 622-626.

30. Doms, A. und Schroeder, M. **GoPubMed: exploring PubMed with the Gene Ontology**. *Nucleic Acids Research 33*, Web Server issue (2005), W783-786.

31. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., u. a. **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Research 32*, Database issue (2004), D258-261.

32. Dietze, H. und Schroeder, M. **GoWeb: a semantic search engine for the life science web**. *BMC Bioinformatics 10 Suppl 10*, (2009), S7.

33. Hoffmann, R. und Valencia, A. **A gene network for navigating the literature**. *Nature Genetics 36*, 7 (2004), 664.

34. Kim, J.-J., Pezik, P., und Rebholz-Schuhmann, D. **MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline**. *Bioinformatics (Oxford, England) 24*, 11 (2008), 1410-1412.

35. Krallinger, M., Izarzugaza, J.M.G., Rodriguez-Penagos, C., und Valencia, A. **Extraction of human kinase mutations from literature, databases and genotyping studies**. *BMC Bioinformatics 10 Suppl 8*, (2009), S1.

36. Yeniterzi, S. und Sezerman, U. **EnzyMiner: automatic identification of protein level mutations and their impact on target enzymes from PubMed abstracts**. *BMC Bioinformatics 10 Suppl 8*, (2009), S2.

37. Brooksbank, C., Cameron, G., und Thornton, J. **The European Bioinformatics Institute's data resources**. *Nucleic Acids Research 38*, Database issue (2010), D17-25.

38. Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., u. a. **DDBJ progress report**. *Nucleic Acids Research 39*, Database issue (2011), D22-27.

39. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., und Sayers, E.W. **GenBank**. *Nucleic Acids Research 38*, Database issue (2010), D46-51.

40. Magrane, M. und Consortium, U. **UniProt Knowledgebase: a hub of integrated protein data**. *Database: The Journal of Biological Databases and Curation 2011*, (2011), bar009.

41. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., und Bourne, P.E. **The Protein Data Bank**. *Nucleic Acids Research 28*, 1 (2000), 235-242.

42. Sigrist, C.J.A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., und Hulo, N. **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic Acids Research 38*, Database issue (2010), D161-166.

43. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., u. a. **BRENDA, the enzyme information system in 2011**. *Nucleic Acids Research*, (2010).

44. Schomburg, D. und Schomburg, I. **Enzyme databases**. *Methods in Molecular Biology (Clifton, N.J.) 609*, (2010), 113-128.

45. Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., und Schomburg, D. **BRENDA, AMENDA and FRENDA: the enzyme information system in 2007**. *Nucleic Acids Research 35*, Database issue (2007), D511-514.

46. Chang, A., Scheer, M., Grote, A., Schomburg, I., und Schomburg, D. **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Research 37*, Database issue (2009), D588-592.

47. Salton, G. ***The SMART Retrieval System - Experiments in Automatic Document Processing***. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

48. Altman, R.B., Bergman, C.M., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., u. a. **Text mining for biology - the way forward: opinions from leading scientists**. *Genome Biology 9*, Suppl 2 (2008), S7.

49. Leopold, E. und Kindermann, J. **Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?** *Mach. Learn. 46*, 1-3 (2002), 423-444.

50. Salton, G., Wong, A., und Yang, C.S. **A vector space model for automatic indexing**. *Commun. ACM 18*, 11 (1975), 613–620.

51. Salton, G. **Term-weighting approaches in automatic text retrieval**. *Information Processing & Management 24*, 5 (1988), 513-523.

52. Joachims, T. **Transductive Inference for Text Classification using Support Vector Machines**. *International Conference on Machine Learning (ICML)*, (1999), 200–209.

53. Cochrane, G., Karsch-Mizrachi, I., und Nakamura, Y. **The International Nucleotide Sequence Database Collaboration**. *Nucleic Acids Research 39*, Database issue (2011), D15-18.

54. Kaminuma, E., Mashima, J., Kodama, Y., Gojobori, T., Ogasawara, O., Okubo, K., Takagi, T., und Nakamura, Y. **DDBJ launches a new archive database with analytical tools for next-generation sequence data**. *Nucleic Acids Research 38*, Database issue (2010), D33-38.

55. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., u. a. **The European Nucleotide Archive**. *Nucleic Acids Research 39*, Database issue (2011), D28-31.

56. Maglott, D., Ostell, J., Pruitt, K.D., und Tatusova, T. **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Research 39*, Database issue (2011), D52-57.

57. Velankar, S. und Kleywegt, G.J. **The Protein Data Bank in Europe (PDBe): bringing structure to biology**. *Acta Crystallographica. Section D, Biological Crystallography 67*, Pt 4 (2011), 324-330.

58. Bodenreider, O. **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Research 32*, Database issue (2004), D267-270.

59. Aronson, A.R. **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program**. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, (2001), 17-21.

60. Hofmann, O. und Schomburg, D. **Concept-based annotation of enzyme classes**. *Bioinformatics 21*, 9 (2005), 2059-2066.

61. Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., und Schomburg, D. **BRENDA: a resource for enzyme data and metabolic information**. *Trends in Biochemical Sciences 27*, 1 (2002), 54-56.

62. Baer, D. ***Duden, Fremdwörterbuch : auf der Grundlage der neuen amtlichen Rechtschreibregeln***. Dudenverl., Mannheim [u.a.], 2001.

63. Jirapongsananuruk, O., Niemela, J.E., Malech, H.L., und Fleisher, T.A. **CYBB mutation analysis in X-linked chronic granulomatous disease**. *Clinical Immunology (Orlando, Fla.) 104*, 1 (2002), 73-76.

64. Scagliotti, G.V., Selvaggi, G., Novello, S., und Hirsch, F.R. **The biology of epidermal growth factor receptor in lung cancer**. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 10*, 12 Pt 2 (2004), 4227s-4232s.

65. Saw, S. und Aw, T.C. **Age-related reference intervals for free and total prostate-specific antigen in a Singaporean population**. *Pathology 32*, 4 (2000), 245-249.

66. Yeh, K.C., Deutsch, P.J., Haddix, H., Hesney, M., Hoagland, V., Ju, W.D., Justice, S.J., Osborne, B., u. a. **Single-dose pharmacokinetics of indinavir and the effect of food**. *Antimicrobial Agents and Chemotherapy 42*, 2 (1998), 332-338.

67. Tabrizi, P., Wang, L., Seeds, N., McComb, J.G., Yamada, S., Griffin, J.H., Carmeliet, P., Weiss, M.H., und Zlokovic, B.V. **Tissue plasminogen activator (tPA) deficiency exacerbates cerebrovascular fibrin deposition and brain injury in a murine stroke model: studies in tPA-deficient mice and wild-type mice on a matched genetic background**. *Arteriosclerosis, Thrombosis, and Vascular Biology 19*, 11 (1999), 2801-2806.

68. Joachims, T. **Making large-Scale SVM Learning Practical**. In, B. Schölkopf, C. Burges und A. Smola, hrsg., *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, 1999, pp. 169–184.

69. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., u. a. **Top 10 algorithms in data mining**. *Knowledge and Information Systems 14*, 1 (2008), 1-37.

70. Matthews, B.W. **Comparison of the predicted and observed secondary structure of T4 phage lysozyme**. *Biochimica Et Biophysica Acta 405*, 2 (1975), 442-451.

71. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., und Nielsen, H. **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics (Oxford, England) 16*, 5 (2000), 412-424.

72. Cohen, J. **Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit**. *Psychological Bulletin 70*, 4 (1968), 213-220.

73. Fawcett, T. **An introduction to ROC analysis**. *Pattern Recognition Letters 27*, 8 (2006), 861-874.

74. **Poppler PDF rendering library**. http://poppler.freedesktop.org.

75. **ROCR: Classifier Visualization in R**. http://rocr.bioinf.mpi-sb.mpg.de.

76. Hunter, L. und Cohen, K.B. **Biomedical language processing: what's beyond PubMed?** *Molecular cell 21*, 5 (2006), 589-594.

77. Lin, J. **Is searching full text more effective than searching abstracts?** *BMC Bioinformatics 10*, (2009), 46.

78. Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C., und Hunter, L.E. **The structural and content aspects of abstracts versus bodies of full text journal articles are different**. *BMC Bioinformatics 11*, (2010), 492.

79. ***Fact SheetMedical Subject Headings (MeSH®)***. http://www.nlm.nih.gov/pubs/factsheets/mesh.html, 2011.

80. Dr. rer . nat. Antje Chang. **Persönliche Kommunikation**. 2011.

81. Krippendorff, K. ***Content Analysis: an Introduction to its Methodology***. Sage Publications, 1980.

82. Landis, J.R. und Koch, G.G. **The measurement of observer agreement for categorical data**. *Biometrics 33*, 1 (1977), 159-174.

83. Kim, J.-D., Ohta, T., Tateisi, Y., und Tsujii, J. **GENIA corpus--semantically annotated corpus for bio-textmining**. *Bioinformatics (Oxford, England) 19 Suppl 1*, (2003), i180-182.

84. Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., und Rebholz-Schuhmann, D. **Assessment of disease named entity recognition on a corpus of annotated sentences**. *BMC Bioinformatics 9 Suppl 3*, (2008), S3.

85. Rebholz-Schuhmann, D., Yepes, A.J.J., Van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., u. a. **CALBC silver standard corpus**. *Journal of Bioinformatics and Computational Biology 8*, 1 (2010), 163-179.

86. Lohmann, S. und Smolenski, A. ***cGMP abhängige Proteinkinasen und ihre Bedeutung für die kardiovaskuläre Funktion***. Bericht zu Drittmittelprojekt SFB 355 TP B4-D, 2001.

87. Garcia-Dorado, D., Agulló, L., Sartorio, C.L., und Ruiz-Meana, M. **Myocardial protection against reperfusion injury: the cGMP pathway**. *Thrombosis and Haemostasis 101*, 4 (2009), 635-642.

88. Chen, M. und Wang, J. **Gaucher disease: review of the literature**. *Archives of Pathology & Laboratory Medicine 132*, 5 (2008), 851-853.

89. Pastores, G.M. **Recombinant glucocerebrosidase (imiglucerase) as a therapy for Gaucher disease**. *BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy 24*, 1 (2010), 41-47.

90. Perrone, M.G., Scilimati, A., Simone, L., und Vitale, P. **Selective COX-1 inhibition: A therapeutic target to be reconsidered**. *Current Medicinal Chemistry 17*, 32 (2010), 3769-3805.

91. Kwon, K.H., Barve, A., Yu, S., Huang, M.-T., und Kong, A.-N.T. **Cancer chemoprevention by phytochemicals: potential molecular targets, biomarkers and animal models**. *Acta Pharmacologica Sinica 28*, 9 (2007), 1409-1421.

92. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., und Hassanali, M. **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Research 36*, Database issue (2008), D901-906.

93. Dr. rer . nat. Marice Scheer. **Persönliche Kommunikation**. 2011.

94. Dipl. bioinf. (FH) Sandra Placzek. **Persönliche Kommunikation**. 2011.

95. Baumgartner, R., Meder, G., Briand, C., Decock, A., D'arcy, A., Hassiepen, U., Morse, R., und Renatus, M. **The crystal structure of caspase-6, a selective effector of axonal degeneration**. *The Biochemical Journal 423*, 3 (2009), 429-439.

96. Creagh, E.M., Conroy, H., und Martin, S.J. **Caspase-activation pathways in apoptosis and immunity**. *Immunological Reviews 193*, (2003), 10-21.

97. Viana, R.J.S., Fonseca, M.B., Ramalho, R.M., Nunes, A.F., und Rodrigues, C.M.P. **Organelle stress sensors and cell death mechanisms in neurodegenerative diseases**. *CNS & Neurological Disorders Drug Targets 9*, 6 (2010), 679-692.

98. Ding, J., Berleant, D., Nettleton, D., und Wurtele, E. **Mining MEDLINE: abstracts, sentences, or phrases?** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, (2002), 326-337.

99. Groves, T. und Abbasi, K. **Screening research papers by reading abstracts**. *BMJ (Clinical Research Ed.) 329*, 7464 (2004), 470-471.

100. von Elm, E. **Writing the abstract: completeness and accuracy matter**. *European Journal of Anaesthesiology 28*, 7 (2011), 483-484.

101. Ward, L.G., Kendrach, M.G., und Price, S.O. **Accuracy of abstracts for original research articles in pharmacy journals**. *The Annals of Pharmacotherapy 38*, 7-8 (2004), 1173-1177.

102. Neumann, E. **A Life Science Semantic Web: Are We There Yet?** *Science's STKE 2005*, 283 (2005), pe22-pe22.

103. Rzhetsky, A., Seringhaus, M., und Gerstein, M. **Seeking a New Biology through Text Mining**. *Cell 134*, 1 (2008), 9-13.

Danksagung

An dieser Stelle möchte ich allen danken, die mir im Verlauf dieser Arbeit hilfreich zu Seite standen. Mein Dank gilt allen Mitgliedern der Arbeitsgruppe Schomburg für die angenehme Arbeitsatmosphäre und die ständige Bereitschaft zum interessierten Gedankenaustausch.

Im Besonderen danke ich Herrn Prof. Dr. Dietmar Schomburg für die interessante Themenstellung und die unterstützende Betreuung. Des Weiteren gilt mein Dank Herrn Dr. Maurice Scheer für die ständige Bereitschaft mit fachlichem Rat und anregenden Ideen meine Arbeit zu begleiten sowie der freundlichen Unterstützung in seiner Funktion als BRENDA Webmaster bei der Integration der Ergebnisse. Darüber hinaus möchte ich mich bei dem Systemadministrator unserer Arbeitsgruppe Adam Podstawka bedanken, der zu allen Zeiten im Einsatz ist um zu helfen, zu reparieren und zu erläutern. Ebenfalls gilt mein Dank Frau Dipl. bionf. (FH) Sandra Placzek für die freundliche Unterstützung und Zusammenarbeit.

Mein besonderer Dank gilt Frau Dr. Antje Chang und Frau Dr. Ida Schomburg, die mir nicht nur stetig durch ihren motivierenden Einfluss und fachlich kompetenten Rat zur Seite standen, sondern darüber hinaus in wichtigen Abschnitten meiner bisherigen akademischen Arbeit immer hilfreich unterstützt haben.

Ebenfalls möchte ich mich für die finanzielle Förderung dieser Arbeit durch die Projekte *Free European Life-Science Information and Computational Services* (FELICS) und *Serving Life-science Information for the Next Generation* (SLING) der europäischen Union bedanken.

Daneben möchte ich mich bei meiner Familie bedanken, die mir in ihrer Gesamtheit ein liebevolles Umfeld und ein unterstützender Hintergrund ist. Vor allem möchte ich mich ganz herzlich bei meinem Mann bedanken, der bei Allem was ich tue, ein liebender Partner, verständnisvoller Freund und beflügelnder Quell der Inspiration ist.